Data analytical stability of measuring brain activation in fMRI studies.

Sanne Roels

Promotor: Prof. dr. Beatrijs Moerkerke
Co-promotor: Prof. dr. Tom Loeys

Proefschrift ingediend tot het behalen van de academische graad van Doctor in de Psychologie

2016

Table of Contents

Acknowledgements

I am grateful to many people for their help during the completion of this dissertation. Please pardon my Dutch in the expression of this gratitude.

Ik ben veel dank verschuldigd aan Bieke en Tom, mijn promotor en co-promotor. In het bijzonder wens ik jullie te bedanken voor de begeleiding en samenwerking. Deze was bij wijlen intens maar steeds met ruimte voor een humoristische verpozing. Wat mij vooral zal bijblijven is het streven naar een blijvend verbeteren en aanscherpen van de invalshoeken bij het opzetten en verwerken van de studies. Ook voor het vele geduld bij het herwerken van de manuscripten wil ik jullie bedanken. Ik geef toe, hierbij durfde ik jullie aanmerkingen wel eens met een pejoratieve term als kipkap omschrijven. Dit stond evenwel niet in verhouding tot de appreciatie ervan! Ik heb alvast het gevoel dat ik veel bijgeleerd heb gedurende de afgelopen jaren. Bieke en Tom, dankjewel om een mentor geweest te zijn!

Next I want to thank the members of my guidance committee – Prof. dr. Daniele Marinazzo, dr. Seppe Santens and Prof. dr. Stefan Van Aelst, for their sound advices during the past 6 years.

I also want to thank all members of the exam jury for taking their time to read and evaluate this dissertation.

Further, only warm and grateful feelings for the people that were involuntarily forced to face me, in the office, on an almost daily basis. Margarita, since you switched the chilly Ghent for a warm French town last year, we no longer share offices. Thank you for the past years and may all go well with your husband and your baby to be! Wouter, steeds bereid om alle discussies naar een hoger niveau te tillen, zowel qua inhoud als volume. Bedankt hiervoor! Han, de immer spitsvondige en humoristische tjoolder van het bureau. Bedankt voor alle grollen! Freya, we deelden het bureau slechts gedurende een korte periode, niettemin een periode vol plezier. Dankjewel hiervoor! It was a true pleasure sharing time with you.

Eveneens ben ik een woord van dank verschuldigd aan de collega's van de vakgroep data-analyse voor de vele interessante discussies en vaak onverwacht absurde wendingen tijdens een lunch of tea break. In het bijzonder richt ik ook een woord van dank aan Isabelle voor het volbrengen van de vele praktische en minder praktische zaken voor en achter de schermen. Dankjewel!

Verder wens ik aan mijn studiegenoten/mede-PhD'ers Joke, Johan,

Kim, Nathalie, Senne en Wout een woord van appreciatie te richten. Joke en Johan, bedankt voor de gezellige tijd tijdens onze gezamenlijke route gedurende de afgelopen tiental jaar, van de expsy over de mastat tot PhD. Joke, de vele interessante discussies waren een groot genoegen. Eveneens dankjewel voor het beschikbaar stellen van de lay-out van dit boek. Veel succes aan de andere kant van de wereld!

Hiernaast wens ik ook mijn vrienden te bedanken voor hun voortdurende vriendschap en warmte, niet alleen doorheen dit doctoraat. Verder wens ik ook Heleen en Wouter te bedanken om hun oog te werpen op delen van dit manuscript.

Ook mijn broer, zussen en ouders wens ik te bedanken voor hun onvoorwaardelijke steun doorheen het leven.

En tot slot, Evelien en Stig. Jullie hebben beiden een levenslustige glimlach die alles in het juiste perspectief brengt. Dankjewel voor alle liefde en steun!

Sanne,
Maart 2016

1 General Introduction

The human brain, that has been fascinating people for centuries, is the central topic of this doctoral dissertation. Although we do not address the *matter* – grey or white – directly, we study how to assess the data analytical stability of brain measurements. More specifically we investigate the stability of statistical conclusions in studies using functional Magnetic Resonance Imaging (fMRI), a dominant tool in the field of cognitive neuroscience nowadays.

With this introduction we aim to set the mind for the 4 main chapters. First, we will present an elective –maybe even an eclectic– historical overview that focuses on 3 milestones in the development of fMRI. Next, we lay out the statistical modeling used in fMRI studies, in a not too technical and non-exhaustive way. These two sections allow a natural introduction for the main subject of this dissertation in Section 1.3. We conclude with the outline of this dissertation.

1.1 Elective Historical Overview

1.1.1 William James' view

William James earned a dominant role in history of psychology and neuroscience with his magnum opus, *The Principles of Psychology*. This homo universalis, with a background in painting, biology, medicine and philosophy, was neither the first in history to describe phenomena that dealt with perception, emotions or thinking, nor was he the first to explicitly link these to the working of the nervous system (Sandrone et al., 2014). However, in his master piece there is one particularly interesting passage that describes the relation between at the one hand the blood flow and blood volume and at the other hand brain functioning (Goodman, 2013; Buxton, 2009).

> We must suppose a very delicate adjustment whereby the cir-

> culation follows the needs of the cerebral activity. Blood very likely may rush to each region of the cortex according as it is most active, but of this we know nothing. I need hardly say that the activity of the nervous matter is the primary phenomenon, and the afflux of blood its secondary consequence.
>
> James (1890), *The Principles of Psychology, chapter III*

His ideas on the relationship between brain and blood were inspired by the work of Angelo Mosso, an Italian neurologist whose experiments could, by current standards, easily be confused with torturing practices (Sandrone et al., 2014; Raichle, 1998). For more than 100 years this relationschip has fascinated researchers. Despite this keen interest, the underlying mechanisms are not yet entirely understood (Hillman, 2014). With the later discoveries of Ogawa et al. (1992), the link between blood flow and brain activation continues to play a critical role in every fMRI study (see below in Section 1.1.3).

For completeness, we note that almost simultaneously with the theoretical work of James, Roy & Sherrington (1890) proposed an extensive work (on dogs) in which they made the following suggestion

> ...to indicate the existence of an automatic mechanism by which the blood-supply of any part of the cerebral tissue is varied in accordance with the activity of the chemical changes which underlie the functional action of that part ...
>
> Roy & Sherrington (1890), *p. 105*

During the same exciting time period, the theory on how neurons function was also developed. Complementary to the idea of James, this theory, with important contributions of e.g. Rajal and Sherrington – who coined the term synapse, focused on the underlying physiology of nerves. This theorising also described the function of the directed electric potentials present in the neuron fibers that could lead to either inhibition or excitation of neighbouring neurons (M. R. Bennett, 1999).

1.1.2 From Resonance to MRI

In 1944 Isidor I. Rabi was awarded the Nobel prize for Physics for his invention of the principle of Nuclear Magnetic Resonance that allows to measure the magnetic properties of atomic nuclei (Nobelprize.org, 2014).

Subsequent simultaneous work of Mathew E. Purcell and Felix Bloch (also both awarded with a Noble prize for Physics in 1952) lead to the precise measurement of the composition of different materials via the application of radio frequency waves (Buxton, 2009, , p. 68-69). In 1973, Paul Lauterbur proposed to deliberately emit these waves in a controlled way to obtain images, i.e. registrations in space (Lauterbur, 1973).

For the description of the principle underlying Magnetic Resonance, we proceed with a simplification identical to Buxton (2009) [1]. Consider a sample, e.g. a human body or a dead salmon, which is put in a large magnetic field (denoted B_0 in Figure 1.1). The strength of this magnetic field is about 100.000 stronger than the magnetic field of the earth[2]. Around the sample, a coil is positioned perpendicularly on the magnetic field. This coil has both a transmitting and receiving function. Note that, due to the strong magnetic field, all particles that form our sample (e.g. hydrogen elements in the human body) get partially aligned with the magnetic field.

In a first phase, the transmitting phase, an oscillating magnetic field is transmitted in the sample via the coil perpendicular on the strong magnetic field. These pulses are in the radio frequency (RF) range and cause a disturbance. It is denoted as the RF pulse (upper panel in Figure 1.1). Although the strength of this pulse is relatively small compared to B_0, it is able to cause a response itself if it is emitted at specific frequencies. Then, in a second phase, the coil has a receiving function for the emitted signals (lower panel in Figure 1.1). These are the signals that form the basis for Magnetic Resonance Imaging (MRI).

With several extensions and better signal processing since Lauturbur's discovery it became possible to obtain 3D images. The widespread distribution of MRI scanners resulted in a growing popularity of this imaging method. Compared to competing imaging techniques such as X-ray computed tomography (CT) scanning, a technique that uses harmful ionizing radiation, and Positron Emission Tomography (PET), a technique that uses invasive radioactive tracers, the non-invasive MRI was considered as a less harmful alternative.

An example of an MRI scan is presented in Figure 1.2. The left panel is acquired using the different magnetic properties of the hydrogen distri-

[1] While we only present a simplification, we refer to the work of Buxton (2009) for an excellent in-depth book on the physical and physiological underlying principles of (f)MRI.

[2] For completeness, the strength of a 3 Tesla (units of magnetic strength, higher indicates stronger magnetic field) scanner is compared with the magnetic field strength of the earth at the equator.

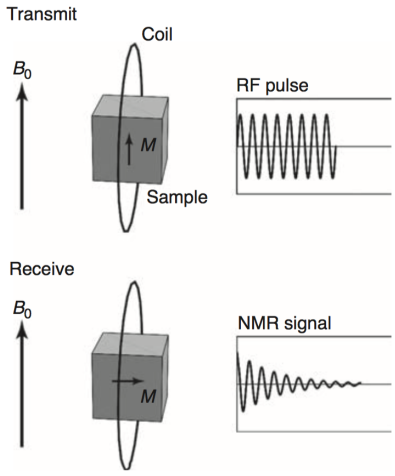


Figure 1.1 The basic Nucleus Magnetic Resonance experiment, fundamental in (f)MRI. (Figure 3.2, p. 71 in Buxton, 2009)

bution in grey and white matter, skull bone and cerebrospinal fluid. As a result, these different structures become visible. In contrast, the image in the right panel of Figure 1.2 is based on different properties which result in a less refined image (see below in Section 1.1.3).

1.1.3 The *functional* in fMRI

Using the MRI scanning protocol, about 100 years after William James wrote about the link between cerebral activity and blood flow, Seiji Ogawa identified different magnetic properties of oxygenated and de-oxygenated blood (Ogawa, Lee, Kat, & Tank, 1990). Using these differences – expressed in different levels of deoxyhemoglobin, the blood flow changes could be detected over time in a non-invasive way. While the first discoveries were in rodents, these were soon extended to humans (see e.g. Kwong, 2012; Ogawa et al., 1992).

In contrast to the already available PET methodology, Ogawa's discovery permitted the detection of blood flow changes in a *non-invasive* way, i.e. without the injection of agents. This signal was denoted as the Blood

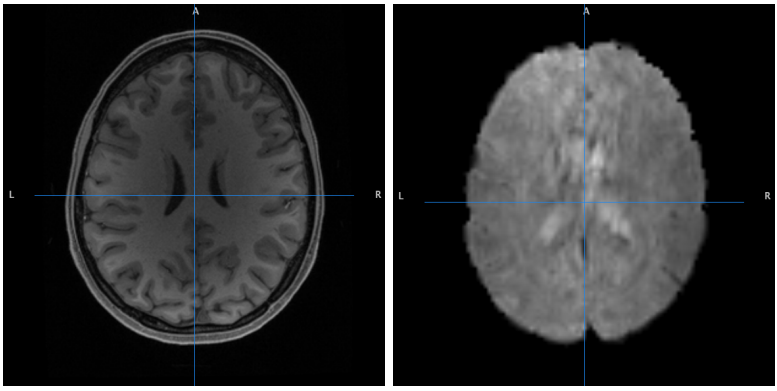


Figure 1.2 Example of a high resolution anatomical scan (left panel) and a Blood Oxygen Level Dependent signal (right panel). Both images are on the same location in the brain and are obtained with the same MRI scanner. The images are obtained from the Human Connectome Project, a large freely available dataset (Van Essen et al., 2012).

Oxygenation Level Dependent (BOLD) contrast and is measured over time. The temporal resolution with wich brain volumes can be scanned over time (i.e. the *TR*: the time to repetition) is bounded by how fast the scanner can register a new volume.

From the two scans in Figure 1.2, the difference in resolution between a functional image (right panel) and an anatomical image (left panel) is apparent. Both the speed of the image acquisition and the properties of the crucial particles (deoxyhemoglobin vs. hydrogen) contribute to the differences.

A simple fMRI experiment A dominant idea in cognitive neuroscience experiments is to verify whether cognitive stimulation leads to specific (and regional) brain activation. Although the experimenters toolbox currently consists out of a wide range of designs (see e.g. Amaro & Barker, 2006), in principle the experiments are variations on one of the original experiments of Ogawa (Ogawa et al., 1992). In this experiment, the difference between the observed BOLD signal in a condition with visual stimulation (light) and the observed BOLD signal in a condition with no stimulation (no light) allowed for the detection of concurring activation changes in the visual cortex. In Section 1.2, we present a second example.

1.1.4 Conclusion: A Well-established Methodology?

The idea that ongoing brain activation is associated with ongoing changes in the cerebral blood flow dates back from the late 19th century. It took however more than a century to demonstrate these changes over time in a non-invasive way. The wide-spread use of the MRI scanners facilitated the use of this methodology in a wide range of fields over the past 25 years. Despite some methodological difficulties, it undoubtably had a massive impact on the view on cognitive neurosciences.

The setup of an fMRI experiment We reconsider three difficulties that affect the setup of an fMRI experiment.

BOLD signal No exhaustive theory that incorporates all aspects of the BOLD signal exists at the moment. Nevertheless, the relationship between regional blood flow changes, changed brain activation and other metabolic processes have been described extensively. Even though these changes co-occur both in spatial and temporal proximity, the scientific community has not yet found a consensus theory (Hillman, 2014).

Donders' rationale The idea of a sensible contrast between a condition with stimulation and a condition without stimulation, a crucial element in most fMRI experiment, dates back to the 19th century work of F. C. Donders. He proposed the following method to measure a thinking process: compare the time it takes to respond to a light cue (A) and compare it to the time it takes to respond to this light cue after having performed a mental task (A+B). By contrasting (A+B) with (A) one is able to compute the time it takes to perform B. One complicating issue here is that in general a non-varying time to perform task B is assumed, i.e. over several repetitions, it takes the same amount of time to perform B (Raichle, 1998).

Methodological difficulties Despite this vast impact on recent scientific evolution, the methodology has some flaws that we summarise here shortly. In comparison with the fast electrical communication between firing neurons, the BOLD signal is slow. Indeed, as James noted, it is without doubt that the BOLD signal is observed after the neuronal activation. Because it follows the neuronal activity, it is inherently an indirect measure of brain activation. Next, even with current scanning settings, the

temporal resolution, or the sampling rate is typically rather slow (i.e. 1-2 seconds).

Impact on cognitive neuroscience The consequences of measuring brain activation in a non-invasive way can not be underestimated for the field of cognitive neurosciences (Rosen & Savoy, 2012). Its non-invasive character, which offered a new and unique window on brain functioning, lead to a boost in fMRI studies on cognitive phenomena. The setup of these experiments allowed to associate brain regions with the execution of a task with more spatial precision than other non-invasive techniques such as electroencephalograms.

In the early days the functional segregation flourished (Poldrack, 2012). This is the localization of regional specialities. A particularly fascinating example of this principle is the *fusiform face area*. This region shows a lot of activation when pictures of faces are shown (Kanwisher, McDermott, & Chun, 1997). Recently, some of the interest shifted towards describing how different brain areas communicate, denoted as functional integration (see e.g. Friston, 2007).

1.2 The Analysis of fMRI Data

The registration of the fMRI data is followed by an extensive statistical analysis. In this dissertation, we focus on the localization of brain activation. We start this section with the outline of a relatively simple experiment and the general underlying concepts in the analysis of fMRI data. Next, we proceed with how to analyse the obtained data. We end this section with how to make conclusions in fMRI studies. For all illustrations in this section we have used the real data of one subject.

An fMRI experiment During this experiment a subject is put in the scanner, and is presented with pictures during 6 periods. In 3 of these periods, the subject sees emotional faces, in the other 3 periods, the subject is presented with neutral pictures (houses). Using the comparison between signals associated with the presentation of houses and with faces, the researcher aims to detect regions specific for the presentation of emotional faces [3] .

[3]This task corresponds to the "Emotion" task in the Human Connectome Data, a large freely available fMRI dataset (Van Essen et al., 2012). The data are taken from 1 subject with identifier 100307 and was minimally pre-processed (Glasser et al.,

The *voxelized* brain The shape of the dataset when it arrives at the researchers computing device is depicted in Figure 1.3. In the left panel the BOLD signal is presented. The signal in the red square (left panel of Figure 1.3) is expressed in numerical values in the right panel of Figure 1.3. It is one large dataset with per time point t ($t : 1 \dots T$, with T the total number of time points) a 3D brain volume, which makes it a large 4D array.

Each of these 3D images is a volume that consists of V voxels, i.e. cubicle volumetric units, with $V = X \times Y \times Z$, with the dimension denoted as (X, Y, Z). For the representation consider a Cartesian coordinate system with an x, y and z axis. As such, each voxel v has a particular coordinate. An example can be found in Figure 1.3 where both images are on *one slice* (on the z-axis) of a 3D brain volume.

Obviously, the larger the dimension of the image, the more precision it can have. A dimension of $(64, 64, 32)$ results in a volume of 131.072 voxels, while a dimension of $(91, 109, 109)$ will results in about 9 times as much voxels to represent the brain image.

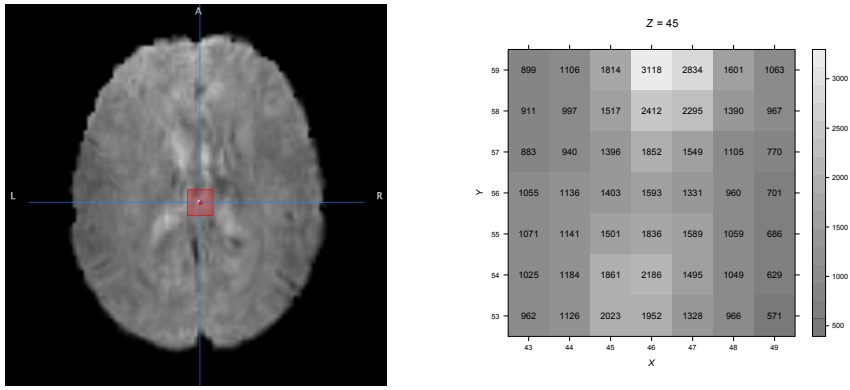


Figure 1.3 The right panel in Figure 1.3 represents the BOLD signal value on one specific time point for the x-coordinates 43-49, for the y-coordinates 53-59 and for the z coordinate 57, this is the zoomed in representation of the BOLD signal in the left panel, which is identical to the right image in Figure 1.2 (image based on: Poldrack et al., 2011, fig. 2.1).

2013) using the FSL software package version 5.0.7 (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012).

Weighing the evidence to localize activation After the data have been collected, it is the aim of the researcher to weigh the evidence. This is typically based on evidence from several subjects. Single subject analyses are however no exception, especially in e.g. pre-surgical diagnosis. For the localization of brain activation, typically, the aim is to obtain a visualization such as these in Figures 1.5b and 1.5c. In these images a test statistic is displayed per voxel. The more intense the color, the more evidence for the voxel being active during the task execution. This type of image is referred to as a statistical parametrical map (SPM).

We will divide the analysis of fMRI data in three parts: 1) the pre-processing of the data; 2) the modeling of the data for one subject and multiple subjects; and 3) how to infer conclusions from the data.

1.2.1 Pre-processing of the Data

The aim of the pre-processing of the data is twofold. At one side, this preparation assures that unwanted noise is removed from the data, but it also prepares the data for next steps in the analysis so that e.g. certain assumptions for inference are satisfied. Note that the order of the following steps, which are typically considered in the pre-processing, depends on the software package (Poldrack et al., 2011).

Slice-timing correction An MRI scanner takes images per plane (e.g. the x and y dimension). The consequence is that only 2 dimensions are registered at a time, and not 3. The result of such sampling of images is that there will be slight time differences between the slices of the volume. Via slice-time correction, this artefact can be (partially) resolved.

Motion correction A second artefact correction is the removal of noise due to head motion. Although most subjects are instructed not to move their head, head motion is almost always inevitable. One important consequence of this motion is a shift in the location of the voxels, i.e. voxel v has a different coordinate over the scans or even within a volume with e.g. different (x, y) coordinates over the z dimension. Via this correction the researcher aims to (partially) resolve this artefact.

Spatial normalization For single subject analyses this pre-processing step is not necessary. However, in multi-subject studies, due to inter-subject variability in the neuro-anatomy of the human brain, it is difficult

to generalize the location information *over* subjects. This is why the individual brain data are transformed to a standardized brain shape. After this transformation, a voxel v has the same location coordinates in all subjects. This allows for a valid aggregation over subjects.

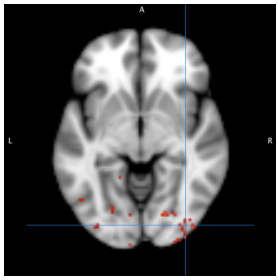
Spatial smoothing This pre-processing step has a drastic impact on the data. The comparison between the left and the right panel in Figure 1.2 makes it clear that the resolution of the BOLD signal is not too high. Through spatial smoothing, the BOLD signal is smeared out over neighbouring voxels via a 3D Gaussian smoothing kernel. These smoothing kernels are described in terms of their width (characterized by the Full Width at the Half Maximum of the kernel (FWHM)). In short, to smooth the signal in a given voxel v on a time point, the weighted average of the signals in all neighbouring voxels is taken. The further voxels are away, the smaller the weight in the average.

While one could consider this as a loss of information, the rationale behind smoothing subsumes the opposite. It is subsumed that smoothing the data will remove irrelevant noise. Furthermore, as the activation typically co-occurs in neighbouring voxels, the activation can be better detected because noise artefacts are cancelled out while activation is accumulated. The second consequence of smoothing is that this allows to use advanced theories to decide whether voxels are activated or not (see below in Section 1.2.3) because it has now larger spatial dependencies.

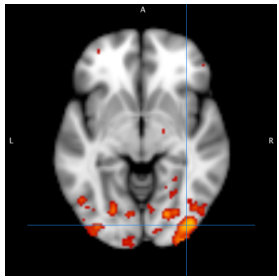
From the examples of spatial smoothing in Figure 1.4, the loss of spatial information becomes clear. Indeed, while no smoothing results in very focal activation spots, more smoothing results in massive activation spots.

1.2.2 Modeling the Data

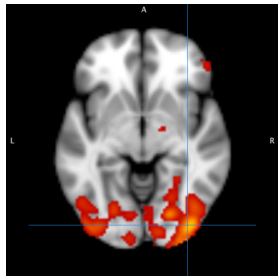
The general linear model In the fMRI literature the general linear model (GLM) is the dominant model to analyze data with (Friston et al., 1995; Lindquist, 2008; Poline & Brett, 2012). In contrast to alternative models which consider the data in an *multi*variate way (e.g. Haxby et al., 2001), the mass-univariate GLM fits a model per voxel v. In what follows we first define the linear model for the first level, i.e. the *within* subject modeling of our toy experiment. Second, we extend this GLM to the situation where multiple subjects are considered.



(a) No smoothing applied.



(b) Gaussian smoothing applied with a kernel width with FWHM of 4mm.



(c) Gaussian smoothing applied with a kernel width with FWHM of 8mm.

Figure 1.4 The effects of spatial smoothing on a SPM that is based on analyses that only differ in the amount of spatial smoothing, expressed in millimeter (mm) FWHM. The more intense the color, the more evidence that the region is involved in the task execution. More smoothing results in more activation in larger regions.

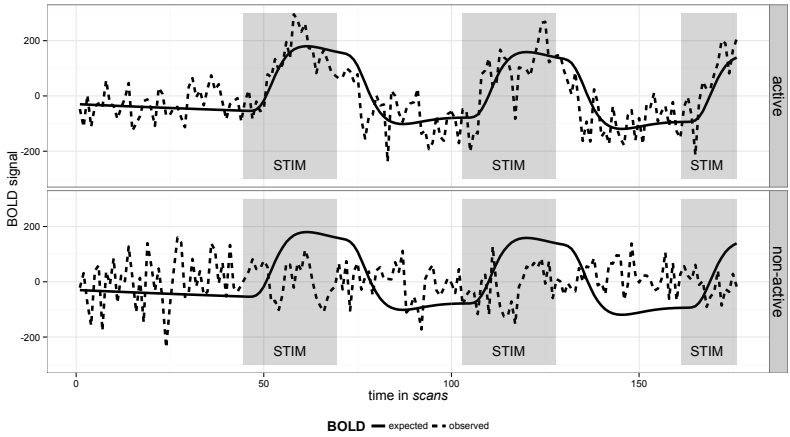
The Model for the *first* Level

The aim of the first level modeling of the fMRI BOLD signal is to correlate the expected BOLD signal under brain activation to the observed signal. The expected signal is constructed using a (pre-)defined shape of the BOLD signal after neural activity. Figure 1.5 illustrates this principle. In Figure 1.5a the two dotted lines represent the observed BOLD signals in 2 voxels, one that will be decided to be active (see Figure 1.5b) and one that will be decided to be non-active (see Figure 1.5c). In these figures, the solid lines represent the expected signal changes under activation while the dotted lines represent the observed signal.

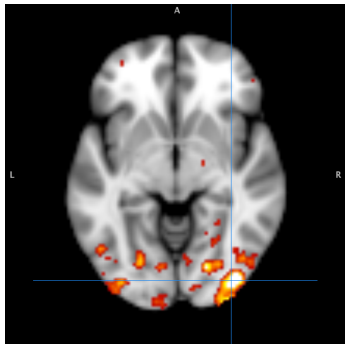
In the simplest case, consider, for a voxel v, the vector with the observed BOLD signals over time $\boldsymbol{y_v}$ with $\boldsymbol{y_v} = (y_{v,1}, \ldots, y_{v,T})$ and with $t : 1 \ldots T$ the time points. Here we will model the model BOLD signal for one stimulus type, i.e. the presentation of fearfull faces. The formulation of the simple linear model is:

$$\boldsymbol{y_v} = \beta_{0,v} + \beta_{1,v}\boldsymbol{x} + \boldsymbol{\epsilon_v}, \tag{1.1}$$

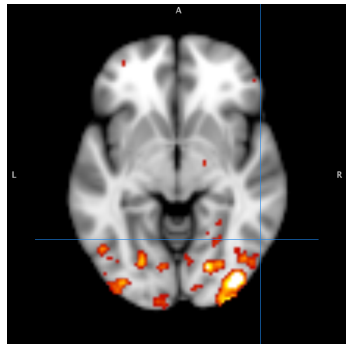
where $\boldsymbol{x} = (x_1, \ldots, x_T)$ represents the values of the expected BOLD signal under brain activation over all time points t, with $\beta_{0,v}$ the intercept (this roughly equals the baseline level), with $\beta_{1,v}$ the effect of the design. In



(a) The expected BOLD signal under activation (full line) plotted against the observed time course of an active voxel (dotted line, upper panel) and against the time course of a non-active voxel (dotted line, lower panel). For the ease of visualization, both time series are centered around 0.



(b) SPM with a crosshair at an active voxel with coordinates $(63, 88, 57)$. The more intense the color, the more evidence for activation.



(c) SPM with a crosshair at an in-active voxel with coordinates $(72, 75, 57)$. The more intense the color, the more evidence for activation.

Figure 1.5 Illustration of the time course of an active and an inactive voxel in 1.5a with a baseline artificially set at 0. In 1.5b and 1.5c the crosshair is at the location of voxel with evidence of being active and with no evidence respectively.

$\boldsymbol{\epsilon_v}$ the noise over the T time points is stored: $\boldsymbol{\epsilon_v} = (\epsilon_{v,1}, \ldots, \epsilon_{v,T})$. In

Figure 1.6, $\boldsymbol{x}$ is plotted in the upper panel.

With the observed data in $\boldsymbol{y_v}$ and $\boldsymbol{x}$ it is possible to estimate β_0 and β_1. If we replace the β_0 and β_1 by their estimates, respectively denoted b_0 and b_1, we can estimate the time course. This is visualized in the lower part of Figure 1.6 for the active and non-active voxel from Figure 1.5a. While b_0 corresponds to the baseline of the signal, b_1 corresponds to the *effect*, i.e. the distance between baseline and the highest level of estimated activation. From Figure 1.5a it is clear that in an active voxel the expected signal closely corresponds to the observed signal, while in a non-active voxel this is not the case (see also the lower panel in Figure 1.6)

The crucial decision on the activation of a voxel v is addressed in detail below in Section 1.2.3.

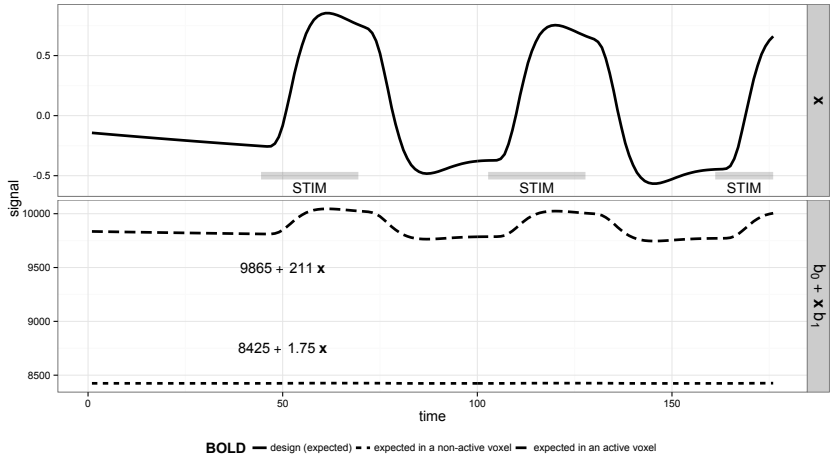


Figure 1.6 Illustration of the time course of an active and an inactive voxel in 1.5a. In 1.5b and 1.5c the crosshair is at the location of voxel with evidence of being active and with no evidence respectively.

The BOLD shape The validity of the model in Equation (1.1) heavily depends on the specification of the expected BOLD signal. It is possible to use very elaborate models for the BOLD signal that are estimated from the data. However, several software packages use a predefined Hemodynamic Responce Function (HRF) that describes the BOLD signal.

In the illustration, we use the popular double-gamma function, i.e. the combination of 2 gamma functions. Using this type of HRF results in an

expected signal pattern as in the upper panel of Figure 1.6. As described above, from this figure it is clear that the BOLD signal is indeed a slow signal. More specifically, we see that the signal changes arise only several scans after the start of the stimulation (indicated with "STIM").

Accounting for the temporal dependencies The BOLD signal is sampled over time. This means that substantial inter-dependencies can exist in the noise over the time points (Woolrich, Ripley, Brady, & Smith, 2001; Worsley et al., 2002). Moreover, these temporal dependencies should be modelled in this first-level modeling. While different modeling strategies exist – with varying success in modeling abilities (Lenoski, Baxter, Karam, Maisog, & Debbins, 2008), most software packages apply a technique called *pre-whitening* if parametric inference is used (see further in Section 1.2.3).

In short, this technique aims at removing the temporal dependencies in the noise. To resolve this, both $\boldsymbol{y_v}$ and $\boldsymbol{x}$ are multiplied with a whitening matrix. Without going into details (which can be found in Chapter 2), if proper whitening occurs this results in more valid conclusions. If the pre-whitening did not occur properly this can impact further conclusions (see e.g. Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012).

Complexity of the design matrix The toy example was introduced as a very simple experiment. While the presentation of only the fearful face could be modelled using Equation (1.1), this simple model can not be used to model the entire experiment. Indeed, to additionally model the presentation of the neutral stimuli, Equation (1.1) is extended as follows. Consider

$$\boldsymbol{y_v} = \beta_{0,v} + \beta_{1,v}\boldsymbol{x_1} + \beta_{2,v}\boldsymbol{x_2} + \boldsymbol{\epsilon_v}, \tag{1.2}$$

where except for $\beta_{2,v}$ and $\boldsymbol{x_2}$ the formulation is identical as above. While $\boldsymbol{x_2}$ represent the expected BOLD signal for the presentation of neutral stimuli, β_2 represents the effect in the signal.

Equations (1.1) and (1.2) can be generalized with:

$$\boldsymbol{y_v} = X\boldsymbol{\beta_v} + \boldsymbol{\epsilon_v}, \tag{1.3}$$

where X is a matrix that contains the expected BOLD signals for the p different conditions in the experimental setup, and $\boldsymbol{\beta_v}$ (with $\boldsymbol{\beta} : (\beta_0, \ldots, \beta_p)$) contains the effects of all conditions. We denote the estimates of $\boldsymbol{\beta}$ with $\boldsymbol{b}$.

Based on the setup of our toy example, one of the research questions is what brain regions are associated with processing emotional information. This can be tested with contrasting the condition with the emotional stimuli and the condition with the neutral stimuli. Such contrast can be modelled neatly with a linear combination of the elements in $\boldsymbol{\beta}$, denoted $\boldsymbol{c\beta}$.

While these contrast can be tested in single subject studies (see below), for the analysis of multiple subjects, it is typically the estimates of the contrast images, with a value per voxel v, denoted as $\boldsymbol{cb}$ that are passed to the second level analysis.

Modeling the *second* Level

The modeling of the second level, i.e. multi-subject analysis, also proceeds via a GLM (e.g. Mumford & Nichols, 2006; Beckmann, Jenkinson, & Smith, 2003).

Consider in a subject m, with $m : 1 \dots M$ and M the total number of subjects in a study, for a given voxel v, $b_{m,v} = \boldsymbol{cb_{m,v}}$, as the input contrast for the group level analysis. $b_{m,v}$ is thus the estimated contrast value per voxel v in subject m based on Equation (1.3) (Beckmann et al., 2003). A linear model, where the index v is dropped for simplicity, is used as follows

$$\boldsymbol{b} = \boldsymbol{X_M}\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{1.4}$$

where $\boldsymbol{b} = b_1, \dots, b_M$, $\boldsymbol{X_M}$ denotes the design matrix which may now contain subject specific covariates, such as gender, age, ... and $\boldsymbol{\eta}$ is the error term with $\boldsymbol{\eta} = (\eta_1, \dots, \eta_M)$.

For the estimation of $\boldsymbol{\gamma}$, the parameters of interest, several procedures are described in the literature (e.g. Beckmann et al., 2003; Mumford & Nichols, 2009; Chen, Saad, Britton, Pine, & Cox, 2013).

1.2.3 Localize Activation in a Multiple Testing Context

In our toy example, the research question is: which brain regions (i.e. which voxels) are associated with the processing of emotional information? To localize brain activation, we need to make a decision on the involvement of brain regions (large regions or voxels). By considering the BOLD signal in a probabilistic context, we can use the null hypothesis significance testing framework. In this framework, to make decisions, we consider the following (simple) heuristic (see e.g. Good, 2005, p. 8):

1. Analyze the problem: identify null hypotheses and the alternative hypotheses.
2. Choose the test statistic.
3. Compute the test statistic.
4. Determine frequency distribution of the test statistic under the null hypothesis.
5. Make a decision with this distribution as a guide.

Based on the previous section, we have analyzed the problem and set up the model for the data.

Setting hypotheses and defining the test statistic

To formalize the decision process, consider the null and alternative hypothesis, respectively denoted H_0 and H_1. Both hypotheses deal with the model parameters $\boldsymbol{\beta}$ (and not the estimates $\boldsymbol{b}$) from Equation (1.3) (see e.g. Casella & Berger, 2002).

Null hypothesis Under the null hypothesis, we state state that there is no association between the presentation of emotional stimuli and the BOLD signal: $H_0 : \ \beta_1 = 0$, i.e. the *effect is zero*. We could also state that there is no difference between the BOLD signals related to emotional stimuli and neutral stimuli: $\beta_1 = \beta_2$ or $\beta_1 - \beta_2 = 0$. In a more general way the H_0 is expressed as $\boldsymbol{c\beta} = \boldsymbol{0}$.

Alternative hypothesis The most general H_1 states that the H_0 is not true, i.e. $H_1 : \ \beta_1 \neq 0$ and $\boldsymbol{c\beta} \neq \boldsymbol{0}$. In a simplified way, this equals saying that the effect is *not* zero. In the fMRI literature, it is however common practice to only look for activation, via $H_1 : \ \beta_1 \geq 0$ and $H_1 : \ \boldsymbol{c\beta} \geq \boldsymbol{0}$.

Test statistic Based on Equation (1.3), consider the parameter estimates, $\boldsymbol{b}$. To construct the test statistic T, we use the the same linear contrast $\boldsymbol{c}$ as in the H_0 to define

$$T = \frac{\boldsymbol{cb}}{\sqrt{\text{var}\,(\boldsymbol{cb})}}, \tag{1.5}$$

where var $(\boldsymbol{cb})$ denotes the variance on the estimated contrast. T is set up analogously for the estimates in Equations (1.1) and (1.4).

Inference

The evidence against H_0 is expressed by T. Conditional on H_0, the higher T, i.e. the larger $\boldsymbol{cb}$ or the smaller $\text{var}(\boldsymbol{cb})$, the more evidence in the direction of H_1.

In a more formal way, this evidence can be derived with Frequency Distribution (FD) of T under H_0. The Probability Density Function (PDF), which is derived from this FD (see below on how to obtain these distributions), permits to determine a p-value. This p-value expresses, conditionally on the H_0, the probability under H_0 to observe a value T at least as large as the *computed* value t:

$$P(T \geq t|H_0)\,. \tag{1.6}$$

The conclusion to either reject H_0 in favor of H_1 or not to reject H_0, is determined by a significance level α. This α denotes to what degree we can incorrectly reject H_0, i.e. to make a wrong decision by incorrectly concluding that a voxel is active (this is also referred to as committing a Type I error, see Table 1.1). In practice, if the p-value is smaller than a pre-defined α one rejects H_0 in favor of H_1. If the p-value is larger or equal to α, one cannot reject H_0.

		Decision	
		Conclude $\mathsf{H_0}$	Conclude $\mathsf{H_1}$
Truth	Active	False Negative (FN) Type II error	*True Positive* (TP)
	Inactive	*True Negative* (TN)	False Positive (FP) Type I error

Table 1.1 Table of events for Null Hypothesis Significance Testing (NHST) in which evidence against a null hypothesis H_0 is evaluated in the direction of an alternative hypothesis H_1.

We distinguish three techniques to guide the conclusion for our research question. In what follows we describe these shortly.

Parametrical inference An often adopted technique to infer conclusions from the data is to use a well-described (or parametrized) FD (and PDF) for T under H_0. The use of such distribution function, e.g. the standard normal distribution function or the Student t distribution, allows for a

fast computation of the p-values. Consider the following simple example to demonstrate how parametrical inference works. The aim is to compare the heights of 10 subjects that are subdivided in 2 groups with labels "A" and "B". In a first step the averages and the standard deviations are calculated in group A and B. In a next step, the test statistic is computed, e.g. a t-test statistic. In the final step, from this t value, the p-value can either be looked in tables or with software.

The validity of these p-values is however bounded by the degree to which the underlying assumptions are complied with. Although the verification of the assumptions is essential, in fMRI data analysis, these assumptions are seldom verified (Luo & Nichols, 2003; Monti, 2011). We come back to this issue below and we further elaborate on assumptions in the section on the correction for multiple testing (cfr. infra).

Permuation-based inference One drawback of parametric inference is that if the underlying assumptions can not be satisfied, the conclusions are poorly guided. Also, not all test statistics have a distribution with a known shape. Permuation-based inference is an empirical alternatives for the FD that relies on *re-shuffling* of group labels. As such, from a collected sample, one can also obtain the FD under H_0. Consider the same example from above for an illustration. First, one computes the statistic (i.e. a function of the observed data but not necessarily a test statistic) for the difference between the two groups in the sample, this is denoted t. Next, all subjects will be reassigned in a random way to either group "A" or group "B". In this way, any systematic relationship between group and height is destroyed. If one now computes this statistic for all the combinations of re-assignments (i.e. $\binom{10}{5} = 252$ in the example), one arrives at an empirically constructed FD under H_0. In a final step t is compared with this empirical distribution to obtain a p-value to guide conclusions.

While such approach is computationally intensive, it has, under the right circumstances, less assumptions to satisfy (see e.g Holmes, Blair, Watson, & Ford, 1996; Winkler, Ridgway, Webster, Smith, & Nichols, 2014, and see also below).

Bootstrap-based inference The bootstrap method (Efron, 1979) is a general method that uses empirically derived samples from one collected

sample[4]. Based on the empirical distribution function one can derive the inferential properties of *a* statistics using the *plug-in* principle (Efron & Tibshirani, 1993, Chapter 4). This empirical distribution function is acquired from the sampling from a sample at hand (with size n) K new samples with size n. The new elements are drawn with replacement and all original observations have equal probabilities $(1/n)$ to get sampled. Consider the example above to illustrate the bootstrap principle. We start with the computation of e.g. the difference between the average heights between the two groups in the sample, and denote it with d. In a second step, we draw *with replacement* 2 new samples of 5 observations from the two original samples (which are denoted with identifiers $A_1, \ldots, A_5$ and $B_1, \ldots, B_5$). This could e.g. result in the following new sample $A_1, A_2, A_2, A_4, A_3, B_1, B_5, B_3, B_4, B_4$. For the plug-in principle, one computes the statistic in all K new samples. From the K samples we derive the empirical distribution of these estimates to guide further inference. Based on the K new samples it is then possible to construct e.g. an interval for d that contains $(1 - \alpha) * 100\%$ of the empirical values. If that interval contains 0, than we cannot reject H_0, if it does not contain 0, than H_0 will be rejected.

The bootstrap method has application under a wide range of models and test statistics (Davison & Hinkley, 1997), making it a flexible alternative for inference. While in permutation and parametric inference more focus is put on the FD under H_0, bootstrap-based inference focuses in general more on the parameters itself (Good, 2005), although properties derived from the FD, such as e.g. p-values, can also be obtained using bootstrapping (see e.g. Davison & Hinkley, 1997). Similar to permutation-based inference, it has, compared to parametric inference less assumptions to satisfy. We further elaborate on this in the next paragraph.

Underlying assumptions From the above, one could incorrectly infer that non-parametric frequency distributions are the cure-all when information is needed to guide decisions. This his however not the case. In what follows, we elaborate on the underlying assumptions that need to be satisfied. The methods are presented from what is typically conceived as less stringent assumptions to more stringent assumptions.

[4]Note however that Good (2005) also describes a parametrical bootstrap. We do not consider such approach here.

Bootstrap-based inference The crucial assumption to bootstrap is that the sample consists of observations that are independently and identically distributed (i.i.d.) (see e.g. Nichols & Hayasaka, 2003, Table 4). I.i.d. denotes that all observations are independent from each other, i.e. the presence of one observation does not affect the probability to observe an other observation. Also, all observations in a sample need to come from the same FD. In contrast to parametric inference, the shape of this distribution is not assumed to be known, although bootstrap procedures exist that assume more a particular shape of the distribution function (see e.g. Davison & Hinkley, 1997; Good, 2005).

If bootstrapping is used in models like Equation (1.4), the validity additionally depends on the underlying assumptions of such model: i.e. the relation between X and y is specified correctly. At last, if this independence of observations cannot be ascertained e.g. due to temporal dependencies, special techniques are advised (see e.g. Lahiri, 2003). These special techniques can rely on additional modeling, which puts in that case additional assumptions on the bootstrap.

Permuation-based inference The valid application of permutation-based inference assumes that the group labels are exchangeable under H_0 next to the independence of observations. This exchangeability implies that under H_0 the joint distribution of the observations is the same for any re-shuffling of the labels (Good, 2005). If this exchangeability cannot be guaranteed, e.g. by the presence of un-modelled nuisance variables that also influence the height, an empirical null hypotheses cannot be created validly and consequent decisions suffer from this. While in situations with a simple (random) group assignment or simple sampling from 2 population (e.g. male/female) exchangeability is a reasonable assumption, in models such as Equation (1.3), more advanced strategies are needed to guarantee the exchangeability (see e.g. Winkler et al., 2014, for an example within neuroimaging studies).

On the other hand, in simple second-level analyses such these in Equation (1.4), when no other confounding variables are modelled, permutation proceeds via *sign-flipping* of the one-column design matrix $\boldsymbol{X_M}$. As such a null hypothesis can be constructed (Nichols & Holmes, 2002). In most cases not all possible "label swaps" are considered, so the observed p-values are most often approximative in nature while exact inference can be attained if all possible permutations are considered (Ernst, 2004).

Parametric inference The valid application of parametric inference is based on strong distributional assumptions, next to the independence of the observations. Indeed, it assumes a fixed and known shape of the FD and PDF. While its simplicity is appealing, carefull verification on the assumptions is necessary. We also illustrate this with our example, which can be seen as a special case of the GLM framework.

In this case parametrical inference puts requirements on the distribution of the residuals ($\boldsymbol{\epsilon}$ in Equation (1.1)).

More specifically, next to the independence of observations, it assumes that $\boldsymbol{\epsilon}$ is normally distributed ($N(0, \sigma_\epsilon)$) and with σ_ϵ known or estimated. Next to these requirements, the relationship between X and y needs to be specified correctly. In the context of our example, we need to have normally distributed residuals in both groups and we need to have the groups specified correctly. If the observations are not independent, e.g. like in the case of fMRI time series, several strategies to deal with repeated measures and clustered data have been developed within the GLM framework (see e.g. Cochrane & Orcutt, 1949; Fitzmaurice, Laird, & Ware, 2004). These strategies impose additional assumption such as a correct specification of the dependency structure to obtain at valid inference.

The multiple testing problem

Testing for activation in a brain volume goes via V simultaneous tests, i.e. one test for every voxel. The resulting amount of false positives over all tests equals αV[5]. More specifically, for a significance level $\alpha = 0.05$ and a volume of $V = 100.000$, this will result in 5.000 spuriously activated voxels when the H_0 in the V voxels, complicating the decision process to determine which voxels are truly active. This excess in False Positives is referred to as the *multiple testing problem*. Given the mass-univariate GLM approach, the dimension of the problem is large and thus particularly problematic.

A dead salmon? This artefact is exotically illustrated in a study on the cognitive abilities of the salmon. Despite the macabre cognitive state of a dead atlantic salmon, in 2011, a group of scientists reported evidence for cognitive abilities in such dead fish (C. M. Bennett, Baird, Miller, & Wolford, 2011). This was however the consequence of not correcting for

[5]For completeness we note that this is under the assumption of independent tests.

multiple testing. Fortunately, several corrections for multiple testing are available.

Clusters? Based on the above, it seems a natural choice to apply a correction of multiple testing on the voxel level, which is indeed a well documented option (e.g. Brett, Penny, & Kiebel, 2007). Nevertheless, uncorrected results might still be reported (Lindquist & Mejia, 2015).

However, the most popular way to address the multiple testing problem is to consider clusters of activation (Woo, Krishnan, & Wager, 2014). Such clusters are random collections of neighbouring voxels that are considered as a whole. The use of such *groups* of voxels reduces the dimension of the multiple testing problem substantially, e.g. from 100.000 to $\sim$ 1.000. Despite this reduction however, on the cluster level one also needs to address the issue of simultaneous testing.

Corrections for multiple testing The aim of the correction for multiple testing, is to control the amount of type I errors (see Table 1.1). The principles underlying these corrections are not specific for the analysis of fMRI and can be applied on both clusters and voxels. With respect to false positives, in the context of multiple tests (i.e. over V simultaneous tests), these errors are reformulated for the Family-wise Error rate control (FWE) and the control on the False Discovery Rate (FDR), two popular approaches to correct for the multiple testing. In a more formal way, these procedures aim to control, using the notation from Table 1.1, respectively:

Familywise Error Rate which is defined as $P(FP \geq 1)$, i.e. the probability to make at least one type I error.

False Discovery Rate which is defined as $E\left(\frac{FP}{FP+TP}\right)$, i.e. the expectation of the proportion of the rejected null hypotheses wich are erroneously rejected (Benjamini & Hochberg, 1995). In the case of no rejections, i.e. $(TP + FP) = 0$, it is set to zero.

These corrections for multiple testing can be applied using parametric, permutation or bootstrap-based inference (see e.g. Nichols & Hayasaka, 2003, for an overview). Consider for example Random Field Theory, this is an approximation that allows to compute FWE corrected p-values in a fast way (Brett et al., 2007). As indicated above, the validity of such parametric approaches is bounded by the degree to wich the assumptions are complied with. One of these assumptions is that the data have to

be sufficiently smooth (Hayasaka & Nichols, 2003). In Section 1.2.1 we have seen that increased smoothing is associated with more distributed activation areas. Fortunately, permutation-based alternatives, for which the data are not subject to the assumption of being sufficiently, have been proposed to correct for multiple testing (Nichols & Hayasaka, 2003).

1.2.4 Conclusion

To select a significant voxel or cluster, the selection mechanism or pipeline in fMRI is fairly challenging. We consider 4 phases in this selection/conclusion procedure. After the setup of the experimental design, one needs to proceed to proper pre-processing, followed by a formal modeling of the design. In the final phase the inferential procedures allow for guided conclusions. It is this collection of choices in the selection procedure that lies between the bare registration of the BOLD signal in the left panel of Figure 1.2 and the SPM's in the lower panel of Figure 1.5.

1.3 Motivation to Assess Data Analytical Stability

Based on the methodological review by Carp (2012), it becomes clear that in principle an infinite amount of combinations of choices can be made to analyze fMRI data. This plethora of combinations in the selection procedure raises the issue of to reliably find activation that is also reproducible. This fundamental issue entails three fundamental questions:

1. How to validly determine activation? Is e.g. a nominal type I error guaranteed?

2. How reliably can the activation be found?

3. With respect to a replication context, how stable is this activation pattern?

In an evaluation of a (new) method, the focus typically lies on the first question. Furthermore, during the recent years, the second question has also been addressed. Indeed, the reliability of activation patterns, i.e. to what degree correspond activation patterns to each other, has received more attention (e.g. C. M. Bennett & Miller, 2010; Specht, Willmes, Shah, & Jäncke, 2003; Gorgolewski, Storkey, Bastin, & Pernet, 2012; Wilke,

2012). In most settings this is defined via test-retest metrics. However, these metrics need identical measures, a strong assumption that is in practice very difficult to verify. Moreover, in such setting it is difficult to make statements about reproducibility since typically only 2 measurements are available. Due to this lack of a (re-)sampling context, it is furthermore difficult to assess the consequences of methodological choices with respect to the stability.

In this dissertation we aim to address the third question with the concept of data analytical stability. A natural definition of stability can be found in the work of Qiu, Xiao, Gordon, & Yakovlev (2006), where the stability of corrections for multiple testing in the the analysis of microarray data is studied. These authors argue that the outcome of a statistical test (or selection procedure) is subject to random fluctuations, and that the result (i.e. the number of differently expressed genes) should be seen as a random variable. One possible cause for this variability is the myriad of methodological choices in the selection procedure.

Due to the selection procedure in fMRI, the set of candidate features (voxels or clusters) is also subject such variation. Stability refers to the ability to replicate the selected features in a replication or (re-)sampling context. We aim to set up the assessment of data analytical stability in fMRI and to demonstrate its use, complementary to the existing questions of validity and reliability. We will thus enable the quantification for a proxy of the reproducibility of the results via the investigation of methodological choices in the pre-processing, the modeling and the inference phase.

Data analytical stability has previously been used to guide strategies for multiple testing in the analysis of micro-array data (e.g. Gordon, Glazko, Qiu, & Yakovlev, 2007). Recently, this has been extended to the analysis of single-subject fMRI analysis (Durnez, Roels, & Moerkerke, 2014). We also further explore the capabilities of data analytical stability in the decision process. In what follows, we outline the structure of this dissertation.

1.3.1 Outline

Chapter 2 An essential part in the assessment of data analytical stability is the (re-)sampling context. We distinguish two possibilities to set up such context: simulations and bootstrap resampling. While the former allows for an in principle unlimited amount of samples, it is limited by the fact that it remains synthetic and consequently possibly not realistic. One

alternative is to bootstrap the data, i.e. resample for the original dataset and create new samples from it. In the first chapter, we investigate several alternatives to bootstrap single subject fMRI data. We focus on the ability to realistically mimic the original data, but we also focus on the inferential properties with different bootstrap scenarios.

Chapter 3 In the second study, we introduce the concept of data analytical stability as an evaluation for choices in cluster-based inference. With a distinction between validity, reliability and reproducibility, we demonstrate the added value of assessing the stability as a proxy for reproducibility. More specifically, we investigate methodological choices for spatial smoothing and for the implementation of cluster-based inference in this study.

Chapter 4 In the third study, we evaluate choices in the second-level analysis of fMRI data using data analytical stability. For this study, we focus on how to aggregate data over subjects and what choices to make for the frequency distribution (parametric versus permutation inference). We additionally investigate the stability of 3 methods to correct for multiple testing.

Chapter 5 In the final study, we use data analytical stability to aid inferential choices in group studies.. Based on results in the third study we elaborate on the added value of stability in the decision process. We demonstrate how the inclusion of data analytical stability in cluster-based inference results in the selection of more activation in a procedure that sets two thresholds. Additionally, we demonstrate how the inclusion of voxel-wise stability measure can improve the interpretation. We extend the cluster-wise results with results for individual voxels within a significant cluster.

Chapter 6 In this chapter we provide a general discussion of the findings in this dissertation and set up a perspective for future research.

Chapter 7 In the final chapter we provide a summary of this dissertation in Dutch.

Chapters 2-5 have been written as stand-alone articles. Consequently, there might be some overlap between these chapters. It should also be

noted that per chapter there is an introduction of the notation, which might not be exactly identical over the chapters. Chapters 2-4 have already been published, Chapter 5 is submitted for publication and full bibliographical details are provided in the respective chapters.

References

Amaro, E., & Barker, G. J. (2006). Study design in fMRI: basic principles. *Brain and cognition*, *60*(3), 220–32.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, *20*(2), 1052–63.

Benjamini, Y., & Hochberg, Y. (1995). *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* (Vol. 57).

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, *1*, 1–5.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–55.

Bennett, M. R. (1999). The early history of the synapse: From plato to sherrington. *Brain Research Bulletin*, *50*(2), 95–118.

Brett, M., Penny, W., & Kiebel, S. (2007). Parametric procedures. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Buxton, R. B. (2009). *Introduction to functional magnetic resonance imaging. principles and techniques.* Cambridge University Press.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Second ed.). Duxbury Advanced Series.

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–90.

Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61.

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their application.* Cambridge: University Press.

Durnez, J., Roels, S., & Moerkerke, B. (2014). Multiple testing in fmri: a case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, *56*(4).

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman Hall.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, *19*(4), 676–685.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* Wiley.

Friston, K. J. (2007). Functional integration. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 36). Elsevier Ltd./Academic Press.

Friston, K. J., Holmes, A., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical Parametric Maps in Functional Imaging: A general Linear Approach. *Human Brain Mapping*, *2*, 189–210.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., ... Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Good, P. (2005). *Permutation, parametric and boostrap tests of hypotheses.* Springer.

Goodman, R. (2013). William james. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2013 ed.). `http://plato.stanford.edu/archives/win2013/entries/james/`.

Gordon, A., Glazko, G., Qiu, X., & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, *1*(1), 179–190.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, *6*, 1–14.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–30.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Hillman, E. M. C. (2014). Coupling Mechanism and Significance of the BOLD Signal: A Status Report. *Annual review of neuroscience*, *37*(1), 161–181.

Holmes, A. P., Blair, R. C., Watson, J. D., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, *16*(1), 7–22.

James, W. (1890). The principles of psychology. In (chap. Chapter III: On Some General Conditions of Brain-Activity.). accessed via `http://psychclassics.yorku.ca/James/Principles/`.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, *17*(11), 4302–11.

Kwong, K. K. (2012). Record of a single fMRI experiment in May of 1991. *NeuroImage*, *62*(2), 610–612.

Lahiri, S. N. (2003). *Resampling methods for dependent data.* Springer-Verlag, Inc.

Lauterbur, P. (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, *242*, 190 - 191.

Lenoski, B., Baxter, L. C., Karam, L. J., Maisog, J., & Debbins, J. (2008). On the performance of autocorrelation estimation algorithms for fmri analysis. *IEEE Journal of Selected Topics in Signal Processing*, *2*, 828-838.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464.

Lindquist, M. A., & Mejia, A. (2015). Zen and the Art of Multiple Comparisons. *Psychosomatic Medicine*, *77*(2), 114–125.

Luo, W.-L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, *19*, 1014-1032.

Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in human neuroscience*, *5*, 28.

Mumford, J. A., & Nichols, T. E. (2006). Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 42–51.

Mumford, J. A., & Nichols, T. E. (2009). Simple group fMRI modeling and inference. *NeuroImage*, *47*(4), 1469–75.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Nobelprize.org. (2014). *Isidor Isaac Rabi - Biographical.* Retrieved from: *http://www.nobelprize.org/nobel_prizes/physics/laureates/1944/rabi-bio.html* on 6 feb 2016.

Ogawa, S., Lee, T., Kat, A., & Tank, D. (1990). Brain magnetic-resonance-imaging with contrast dependent on blood oxygenationbrain magnetic-resonance-imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(24), 9868-9872.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(13), 5951–5955.

Poldrack, R. A. (2012). The future of fMRI in cognitive neuroscience. *NeuroImage*, *62*(2), 1216–1220.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional mri data analysis.* New York: Cambridge University Press.

Poline, J.-B., & Brett, M. (2012). The general linear model and fMRI: Does love last forever? *NeuroImage*, *62*(2), 871–880.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, *7*.

Raichle, M. E. (1998). Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 765–772.

Rosen, B. R., & Savoy, R. L. (2012). FMRI at 20: Has it changed the world? *NeuroImage*, *62*(2), 1316–1324.

Roy, C. S., & Sherrington, C. (1890). On the regulation of the Blood-Supply of The Brain. *The Journal of Physiology*, *11*, 85–158.

Sandrone, S., Bacigaluppi, M., Galloni, M. R., Cappa, S. F., Moro, A., Catani, M., ... Martino, G. (2014). Weighing brain activity with the balance: Angelo Mosso's original manuscripts come to light. *Brain*, *137*(2), 621–633.

Specht, K., Willmes, K., Shah, N. J., & Jäncke, L. (2003). Assessment of reliability in functional imaging studies. *Journal of Magnetic Resonance Imaging*, *17*(4), 463–71.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231.

Wilke, M. (2012). An iterative jackknife approach for assessing reliability and power of FMRI group analyses. *PloS one*, *7*(4), e35578.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–97.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, *91*, 412–9.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–86.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*(1), 1–15.

2 Bootstrapping fMRI Data: Dealing with Misspecification

Abstract The validity of inference based on the General Linear Model (GLM) for the analysis of functional magnetic resonance imaging (fMRI) time series has recently been questioned. Bootstrap procedures that partially avoid modeling assumptions may offer a welcome solution. We empirically compare two voxel-wise GLM-based bootstrap approaches: a semi-parametric approach, relying solely on a model for the expected signal; and a fully parametric bootstrap approach, requiring an additional parameterization of the temporal structure. While the fully parametric approach assumes independent whitened residuals, the semi-parametric approach relies on independent blocks of residuals. The evaluation is based on inferential properties and the potential to reproduce important data characteristics. Different noise structures and data-generating mechanisms for the signal are simulated. When the model for the noise and expected signal is correct, we find that the fully parametric approach works well, with respect to both inference and reproduction of data characteristics. However, in the presence of misspecification, the fully parametric approach can be improved with additional blocking. The semi-parametric approach performs worse than the (fully) parametric approach with respect to inference but achieves comparable results as the parametric approach with additional blocking with respect to image reproducibility. We demonstrate that when the expected signal is incorrect GLM-based bootstrapping can overcome the poor performance of classical (non-bootstrap) parametric inference. We illustrate both approaches on a study exploring the neural representation of object representation in the visual pathway.

This chapter has been published in NeuroInformatics.
Roels, S. P., Moerkerke, B., & Loeys, T. (2015). Bootstrapping fMRI Data: Dealing with Misspecification. *NeuroInformatics, 13,* 337–352.

2.1 Introduction

Functional magnetic resonance imaging (fMRI) is a popular technique in cognitive neurosciences. Its aim is to detect brain regions that are activated during a particular task by looking for task-induced Blood Oxygen Level Dependent (BOLD) signal changes. This is typically done via a massive General Linear Model (GLM) approach (Lindquist, 2008; Carp, 2012).

In the GLM approach, the time course of the measured BOLD signal of each voxel is modeled as a linear combination of different signal components. This allows to test for specific parameters and contrasts to evaluate evidence for activation. The end product consists of a three-dimensional, statistical parametric map (SPM) which graphically displays the test statistics or the p-values corresponding to the voxels in the brain. Thresholding this SPM enables to determine which voxels or which regions in the brain become active during a specific task.

These GLMs rely on several modeling assumptions, including modeling of the time-dependency of consecutive measures for each voxel due to repeated scans, and the specification of a model for the expected BOLD signal. So far, limited research has focused on evaluating the extent to which these assumptions are fulfilled and the impact of violation of assumptions (with a few notable exceptions Luo & Nichols, 2003; Razavi et al., 2003; Zhang, Luo, & Nichols, 2006) but findings suggest that violation makes the GLM-procedure vulnerable to wrong inference (Aguirre, Zarahn, & D'esposito, 1998; Lindquist, Meng Loh, Atlas, & Wager, 2009; Monti, 2011)

A common approach is to model temporal dependency in fMRI-noise with an autocorrelation structure of order 1, an $AR(1)$ structure, common to all voxels. This fairly simple model is not without criticism however and has been shown to insufficiently capture the full extent of the temporal complexity (Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012; Lenoski, Baxter, Karam, Maisog, & Debbins, 2008). Despite the growing consensus that the assumed $AR(1)$ structure might not be optimal, an $AR(1)$ noise model continues to be used in several standard software programs for fMRI data analysis, such as SPM (Glaser & Friston, 2007). Furthermore, several studies regarding the shape of the BOLD signal indicate that the most popular model, the *double gamma* function (Henson & Friston, 2007), is outperformed by other more flexible models for the mean signal under activation (Grinband, Wager, Lindquist,

Ferrera, & Hirsch, 2008; Lindquist et al., 2009). Nevertheless, the GLM approach using these models for noise and expected signal is continued to be used as it provides a relatively simple but powerful model for the detection of signal changes in fMRI research (see e.g. Thyreau et al., 2012).

Bootstrap procedures may offer a solution to conduct valid inference, simultaneously using the strengths of the linear model and taking into account the complex temporal dependencies present in the data (Lahiri, 2003; Davison & Hinkley, 1997). Briefly, for N independent observations, bootstrapping consists of drawing K random samples of length N with replacement from the original set of observations. Inference can then be conducted using the empirical distribution of the parameters of interest over the K bootstrap samples. This bootstrap principle enables valid inference for a whole range of parameters (Davison & Hinkley, 1997). When observations are not independent however, as in fMRI time series, adjustments are needed.

In the fMRI literature on bootstrap procedures two approaches can be distinguished: 1) a GLM-based approach and 2) a signal-decomposition approach. Both aim at obtaining independent chunks of information to bootstrap from. The first approach is based on resampling GLM residuals after temporal decorrelation assuming a parametric form of the temporal variance-covariance structure (Friman & Westin, 2005). The second approach uses an orthogonal decomposition of the BOLD signal based on e.g. wavelet (e.g. Bullmore et al., 2004; Tang, Woodward, & Schucany, 2008) or Fourier transformation (e.g. Friman & Westin, 2005; Laird, Rogers, & Meyerand, 2004). From the comparison of both approaches the parametric GLM-based approach was found to work best in a range of situations (Friman & Westin, 2005). Nonetheless, in the latter study, the parametric structures of both noise and expected signal were assumed to be correct, which may result in overtly optimistic conclusions. Using more complex fractional Brownian drift noise, Tang et al. (2008) further investigated the size of the test for activation based on wavelet decomposition and found this procedure insufficient as inference method (see also Laird et al., 2004).

Moreover, Friman & Westin (2005) relied on a parameterization of the temporal noise structure to explore bootstrapping in the GLM-framework but this is not necessary per se. Indeed, inference based on such approach may suffer from the same weaknesses as the GLM-approach that directly models the variance-covariance structure. Taking into account the dependency by blocking groups of consecutive observations (hereafter referred

to as the semi-parametric approach) may also result in valid inference (Davison & Hinkley, 1997; Lahiri, 2003), but excludes the need for the explicit parametrization of a complex noise structure.

In this paper, we compare a *fully parametric bootstrap approach* and a *semi-parametric bootstrap approach*. The former relies on a parametric formulation of both the mean and the temporal variance-covariance to obtain independent residuals while the latter only relies on a parametric formulation of the mean. We also investigate a parametric approach in which the fully parametric approach is extended by bootstrapping from blocks of whitened residuals. As fMRI data are characterized by a complex set of noise sources (Greve, Brown, Mueller, Glover, & Liu, 2012) and the complexity of the simulated noise can impact the evaluation (Welvaert & Rosseel, 2012), we apply several sources of noise in our simulation. This enables to study the impact of a misspecified noise structure on the performance of both procedures. Additionally, we also investigate the impact of an incorrectly specified mean structure.

The remainder of this paper is structured as follows. In the Method section more details are provided on both bootstrap methods and the notation used in this paper. The performance of both approaches is compared in the Simulation Study section. Applicability of the proposed approach on real data is illustrated in the Real Data Example section. We end with a discussion.

2.2 Method

2.2.1 Modeling and Inference for fMRI Data

In the GLM approach for fMRI data, the observed BOLD signal for each voxel is related to the experimental design. Let $\mathbf{y_i} = (y_{i1}, \ldots, y_{iT})$ represent the BOLD signal for voxel i that is measured on T time points $(t = 1, \ldots, T)$. In total, there are $N = x \times y \times z$ voxels with x, y and z the respective dimensions of the image $(i = 1, \ldots, N)$. The p columns of the design matrix $\mathbf{X}$ reflect a stimulus function for the experimental design (see e.g. Lindquist, 2008). To account for the delay between the onset of the stimulus presentation and the BOLD response, the stimulus function (i.e. the input) is typically convolved with a hemodynamic response function (HRF). Hence, the original stimulus function changes into a convoluted variable in the design matrix. The upper panel of Figure 2.1 represents the scans at which a specific condition under investigation is

either on or off. The convolution of this stimulus function with the HRF is depicted in the middle left panel of Figure 2.1. The most commonly used HRF is a sum of two gamma distributions (i.e. the double gamma function: Henson & Friston, 2007). The model of interest for each voxel is then as follows:

$$\mathbf{y_i} = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \tag{2.1}$$

For each voxel i, $\boldsymbol{\beta}_i$ represents the vector with the regression coefficients and $\boldsymbol{\varepsilon}_i$ denotes the residual error term. Specific contrasts of these coefficients allow to test for activation that is associated with the task or condition under investigation. When fitting model (2.1) one needs to account for the residual correlation between consecutive time points. Let $V\sigma_\varepsilon^2$ denote the assumed variance-covariance matrix of $\boldsymbol{\varepsilon}_i$, with an $AR(1)$ structure as a default choice for V. A matrix Σ_d is then constructed such that

$$\Sigma_d V \Sigma_d^t = \mathbf{I} \tag{2.2}$$

holds. $\mathbf{y_i}$ and $\boldsymbol{X}$ in model (2.1) are consequently post-multiplied (or *whitened*) with matrix Σ_d resulting in the following model:

$$\tilde{\mathbf{y}}_\mathbf{i} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}_i + \tilde{\boldsymbol{\varepsilon}}_i \tag{2.3}$$

If V is correctly specified, $\tilde{\boldsymbol{\varepsilon}}_\boldsymbol{i} \sim N(0, I\sigma_\varepsilon^2)$ and $\tilde{\boldsymbol{\beta}}_i$ can simply be estimated through the Ordinary Least Squares approach (OLS, see e.g. Lindquist, 2008; Cochrane & Orcutt, 1949; Kutner, Nachtsheim, Neter, & Li, 2005). The validity of inference for $\tilde{\boldsymbol{\beta}}_i$ will heavily depend on the degree to which the assumed variance-covariance structure holds (i.e., whether the appropriate whitening has been done) and to what extent the model of the hemodynamic response holds.

2.2.2 Bootstrap for fMRI Data: a Fully Parametric and a Semi-Parametric Approach

Classical bootstrapping consists of repeatedly drawing a set of independent samples out of N independent observations (Lahiri, 2003). When observations are not independent, such as the residuals in model (2.1), the N observations can be grouped into blocks of length ℓ such that the different blocks represent independent subgroups and can be bootstrapped

from. We refer to this approach as a semi-parametric approach as only the expected signal is modeled via $E(\mathbf{y_i})$. Alternatively, one can sample N observations from the whitened residuals of model (2.3) which is a fully parametric approach since a model is imposed on both the expected signal and noise. In the remainder of this section, we further discuss these two approaches and focus on how they deal with the dependency structure V (autocorrelation) in the residuals.

Fully parametric approach

Using the equations (2.2) and (2.3), the approach consists of the following steps for each voxel $i = 1 \dots N$:

1. Estimate $V\sigma_\varepsilon$ under the assumed dependency structure.
2. Determine $\hat{\Sigma}_d$ such that $\hat{\Sigma}_d \hat{V} \hat{\Sigma}_d^t = \mathbf{I}$ holds.
3. Construct $\tilde{\mathbf{e}}_\mathbf{i} = \mathbf{y_i} - \tilde{\mathbf{X}}\hat{\tilde{\boldsymbol{\beta}}}$.
4. For each voxel i draw K samples with replacement from $\tilde{\mathbf{e}}_\mathbf{i} = (\tilde{e}_{i,1} \dots \tilde{e}_{i,T})$ (assumed to be independent) resulting in K samples $\tilde{\mathbf{e}}_\mathbf{i}^\mathbf{k}$ $(k = 1 \dots K)$ with length T.
5. Add the uncorrelated residuals upon the expected signal and recorrelate via ${\mathbf{y_i}}^k = \hat{\Sigma}_d^{-1} \left(\tilde{\mathbf{X}}\hat{\tilde{\boldsymbol{\beta}}} + \tilde{\mathbf{e}}_\mathbf{i}^\mathbf{k} \right)$.
6. Calculate for each sample k the parameter(s) of interest as function of the data $G_T({\mathbf{y_i}}^k)$.
7. Inference is based on $F_K(G_T({\mathbf{y_i}}^k))$ with $F_K(\cdot)$ the empirical distribution obtained through bootstrapping.

Semi-parametric approach

In this approach blocked bootstrapping is applied, i.e. resampling in blocks of ℓ consecutive observations. Blocks are constructed such that the original dependency structure within blocks is retained while the dependency between the blocks is minimal.

In the bootstrap literature several implementations of blocked bootstrap procedures are proposed (for an excellent overview, see Lahiri, 2003). In this study, we opt for the moving block bootstrap algorithm[1] in which

[1] In a pilot study we looked at alternatives such as e.g. circular bootstrap, but no remarkable differences were observed. Results are not shown.

a time series of length T is split into $T-\ell+1$ overlapping blocks of length ℓ. Observation 1 to ℓ will be block 1, observation 2 to $\ell+1$ will be block 2 etc. Then from these $T-\ell+1$ blocks, T/ℓ blocks are randomly drawn with replacement. The block length ℓ needs to be chosen such that observations more than ℓ time units apart are nearly independent.

Using model (2.1), the following steps are followed for each voxel $i = 1 \dots N$:

1. Choose an appropriate block size ℓ (optimal choices will be discussed later).

2. Construct $\mathbf{e_i} = \mathbf{y_i} - \mathbf{X}\hat{\boldsymbol{\beta}}$, with $\hat{\beta}$ obtained from OLS

3. Create $T-\ell+1$ blocks of length ℓ : $\big((e_{i,1}, \dots, e_{i,\ell}), (e_{i,2}, \dots, e_{i,\ell+1}), \dots, (e_{i,T-\ell+1}, e_T)\big)$ and resample with replacement from those block to obtain $\mathbf{e_i}^k$.

4. For each bootstrap sample k, add the bootstrapped residuals to the predicted signal and create $\mathbf{y_i}^k = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e_i}^k$.

5. Calculate for each sample k the parameter(s) of interest as function of the data $G_T(\mathbf{y_i}^k)$.

6. Inference is based on $F_K(G_T(\mathbf{y_i}^k))$ with $F_K(\cdot)$ the empirical distribution obtained through bootstrapping.

The choice of ℓ is crucial in the semi-parametric bootstrap approach. For the moving block bootstrap procedure Politis & White (2004) developed a formal procedure to estimate an optimal block length, ℓ_{opt} . The algorithm is based on a non-parametric estimate of the spectral properties that results in an estimate of the Mean Squared Error (MSE) for a giving block length ℓ. By minimizing the MSE ℓ_{opt} is determined.

Some authors (see e.g. Davison & Hinkley, 1997, p. 391) suggest to combine the fully parametric with blocked bootstrapping (i.e., applying blocked bootstrap after whitening) to gain robustness against misspecification of V. Step 3 of the fully parametric approach by choosing a block length $\ell > 1$ is then modified as follows:

3. Draw K samples with replacement from blocks of length ℓ from $\tilde{\mathbf{e}}_\mathbf{i}$.

This approach will be referred to as a parametric approach with block length ℓ.

Finally, note that we use standardized residuals to reduce bias due to possible non-zero mean of the residuals and to overcome possible homoscedasticity violations (Davison & Hinkley, 1997). In the fully parametric approach, these standardized residuals $\tilde{r}_t$ are defined as follows for each voxel i (we dropped index i):

$$\tilde{r}_t = \frac{\tilde{e}_t - \bar{\tilde{e}}}{\left(\sqrt{1 - \tilde{h}_{tt}}\right)} \tag{2.4}$$

with $t = 1, \ldots, T$ and with h_{tt} the t^{th} diagonal element of $\Sigma_d X \left((\Sigma_d X)^T \Sigma_d X \right)^{-1} (\Sigma_d X)^T$. In the semi-parametric approach the residuals r_t are defined as

$$r_t = \frac{e_t - \bar{e}}{\left(\sqrt{1 - h_{tt}}\right)} \tag{2.5}$$

with $t = 1, \ldots, T$ and with h_{tt} the t^{th} diagonal element of $X(X^T X)^{-1} X^T$.

2.3 Simulation Study

2.3.1 Data Generation

Model for the mean signal

We simulate a 2D volume with 400 voxels over 360 time points using a time to repetition (TR) of 2 seconds. Activation in one brain region is induced by linking the BOLD signal of the voxels to a simple event-related design. In total, 5% of the voxels lie in an activated brain region while the remaining voxels are non-active. BOLD signal over time for each voxel is generated as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \tag{2.6}$$

with $i = 1, \ldots, 400$ and $t = 1, \ldots, 360$. The regressor x_{it} is a function of the design. For the activated voxels, $\beta_1 \neq 0$ (β_1 varies uniformly from 2.1 to 3), while $\beta_1 = 0$ for the non-active voxels.

The regressor in model (2.6) is created by convoluting the simple 0–1 stimulus function corresponding to the event-related design with a HRF. We consider two different HRFs: the canonical double gamma HRF (Henson & Friston, 2007; Glover, 1999) and the HRF obtained by the physio-

logically based *balloon* model (Buxton, Uludağ, Dubowitz, & Liu, 2004). Both are generated using the implementation of the `R` (R Core Team, 2013) library `neuRosim` (Welvaert, Durnez, Moerkerke, Verdoolaeghe, & Rosseel, 2011).

To study the power, we additionally simulate a data set with all 400 voxels activated with $\beta_1 = 0.25$.

Model for the temporal noise

We impose either a simple or a more complex structure on the temporal correlation of noise component $\boldsymbol{\varepsilon}$. For the simple noise structure, we consider respectively a stationary $AR(1)$ and $AR(2)$ model. For the $AR(1)$ structure, the noise component is generated as follows (Chatfield, 2000):

$$\varepsilon_t = \xi_t + \phi\varepsilon_{t-1} \tag{2.7}$$

with $\xi_t \sim N(0, \sigma_\xi^2)$. Note that the voxel index i is dropped to simplify notation. Provided that $|\phi| < 1$, ε_t has a finite variance and results in a stationary time series with

$$E(\varepsilon_t) = 0 \text{ and } \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2 = \frac{\sigma_\xi^2}{1-\phi^2}.$$

Under this model, temporal dependence between measures at time t and time $t-k$ is as follows: $\rho_{t,t-k} = \phi^{|k|}$ with $k = 0, \ldots, T$ and $t - k \geq 1$. In our simulations, we set the first-order temporal dependence ϕ equal to 0.35.

Noise under the $AR(2)$ structure is generated as follows:

$$\varepsilon_t = \xi_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} \tag{2.8}$$

with $\xi_t \sim N(0, \sigma_\xi^2)$. Provided that $\phi_1 + \phi_2 < 1$ and $\phi_1 - \phi_2 > -1$ and $\phi_2 > -1$ holds, and thus stationarity is achieved, the following properties hold:

$$E(\varepsilon_t) = 0 \text{ and } \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2 = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma_\xi^2}{[(1-\phi_2)^2 - \phi_1^2]}.$$

First and second-order temporal dependencies are as follows: $\rho_{t,t-1} = \frac{\phi_1}{1-\phi_2}$ and $\rho_{t,t-2} = \frac{\phi_1^2}{1-\phi_2} + \phi_2$. In the simulation study, we set ϕ_1 and ϕ_2 equal to 0.3 and 0.15.

For the complex noise scenario, long-term noise is added. More specifically, Fractional Brownian Motion (Tang et al., 2008; Gudbjartsson & Patz, 2005) is added to the $AR(1)$ noise with the same ϕ, resembling a drift-like component that can be observed in fMRI data (see e.g. Lazar, 2008; Greve et al., 2012). Furthermore, ξ_t in model (2.7) was sampled from a Rice distribution:

$$f(\xi_t) = \frac{\xi_t}{\sigma^2} exp\left(-\frac{\xi_t^2 + |v|^2}{2\sigma^2}\right) I_0\left(\frac{\xi_t |v|)}{\sigma^2}\right) \text{ if } \xi_t > 0 \tag{2.9}$$

with $I_0(\xi_t)$ a modified Bessel function of the first kind with order zero. We choose σ and v equal to 1. Through rescaling, we assure that this component does not cause a difference on the signal that exceeds 3% of the baseline level. Examples of different time courses of the signal for active and inactive voxels for the simulation study are shown in the lower panel of Figure 2.1.

Finally note that although spatial correlation is inherent to fMRI data, we do not impose it here in our simulation study. As in Friman & Westin (2005), we focus on bootstrapping from fMRI time series and the effect of temporal correlation on resulting procedures. Since the GLM-approach for fMRI-analysis is typically performed in a voxel-wise manner, the absence of spatial correlation has no impact on the conclusions of voxel-wise measures. However, we will inspect the spatial variability (see further) in the evaluation.

2.3.2 Modeling and Bootstrapping

For each simulated volume, data are analyzed with model (2.1) using the standard double gamma HRF. This corresponds to a correctly specified mean structure when the double gamma HRF is used for data generation and in a misspecified mean model when the balloon model is used to simulate the data. For the fully parametric bootstrap procedure, we use the whitening algorithm as implemented in SPM8 (Wellcome Trust Centre for Neuroimaging U.C.L, 2010). More specifically, the temporal dependency structure is globally estimated through one matrix Σ_d for all voxels of the original volume. By default SPM8 uses a model for V which very closely resembles $AR(1)$, consisting out of autocorrelation part and a noise part (see manual SPM8: Wellcome Trust Centre for Neuroimaging U.C.L, 2010).

We additionally implement a variable $AR(1)$ structure among voxels.

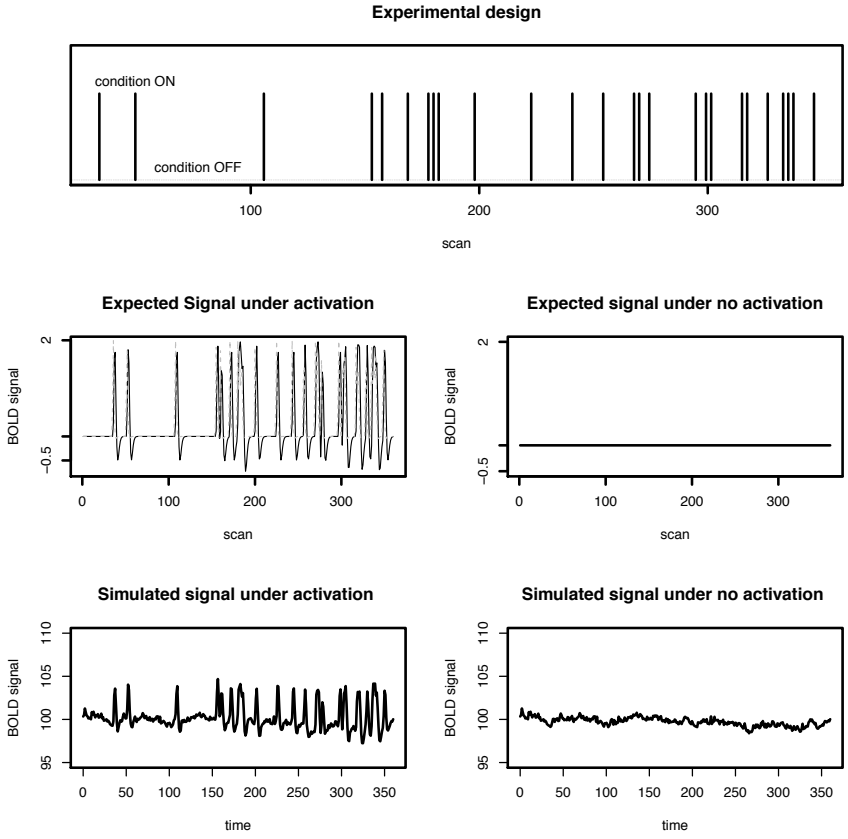


Figure 2.1 Upper panel: the stimulus activation function of the event-related design; middle panels: the convolution of the stimulus activation function with the a) the *double gamma* (black) and b) the *balloon* (grey) HRF which is the expected or mean signal ($X\beta$) under activation (left) and under no activation (right); bottom panels: the simulated signal under the complex noise structure for activated voxels (left) and voxels that were inactive (right). An $AR(1)$ Rician noise model with $\phi = 0.3$ is depicted.

This implementation is characterized by the determination of $V\sigma_\varepsilon$ per voxel i rather than globally, making it more computationally more intensive. Bootstrapping is then applied on the 'whitened' residuals.

For the semi-parametric bootstrap procedure, no whitening is applied first. Both for the parametric and semi-parametric bootstrap approach,

we next consider block lengths ℓ varying from rather short (2) to longer (5, 20 and 40).

Note that when $\ell = 1$ in the fully parametric bootstrap, the corresponding procedure is sometimes referred to as *whitening bootstrap*. $\ell = 1$ in the semi-parametric approach corresponds to the *i.i.d. bootstrap* as it ignores all temporal correlation. K is set to 500 bootstrap samples.

In the semi-parametric bootstrap, we additionally use the method of Politis & White (2004) (an implementation in the R-package `np`: Hayfield & Racine, 2008) to choose an optimal block length ℓ_{opt} for each voxel i.

2.3.3 Evaluation Criteria

To evaluate the performance of the different bootstrap approaches for drawing inference, we assess the nominal type I error rate and power of the test for activation. As bootstrap samples represent replicates of the original data, we also investigate to what extent the bootstrap samples resemble the original images.

Inferential properties

Test size To yield valid inference for the test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ in model (2.6), the observed test size (i.e. the false positive rate) of the bootstrap procedures under consideration should equal the nominal test size α. More specifically, we assess for each voxel whether the bootstrap confidence interval formed by the $\alpha/2$ and the $1 - \alpha/2$ quantile of $F_K(b_1^k)$ contains zero or not. The proportion of false positives averaged over all non-active voxels is then used to determine the empirical test size.

Power The power is determined as the number of rejections of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ in model (2.6) when $\beta_1 = 0.25$. Again, we use bootstrap confidence intervals for this purpose. The proportion of correct rejections is averaged over all active voxels in the volume and is then used to determine the empirical power.

Image reproducibility

Because time series of images are difficult to compare we rely on the distributional properties of summary statistics per time series. We assess the reconstruction of the temporal correlation between two consecutive time

points (i.e., the first-order temporal correlation $\rho_{(t,t-1)}$) and the temporal correlation between two time points with a lag of two (i.e., the second order temporal correlation $\rho_{(t,t-2)}$) and the variance over time points per voxel (σ_t^2). We also study the in-scan variability (σ_s^2) of the bootstrap images.

These quantities are compared to the observed value in the original volume in a way similar to the method used by Bellec, Perlbarg, & Evans (2009). More specifically, the median value of each of these characteristics is repeatedly calculated over the $K = 500$ bootstrap volumes. For each measure, a smooth density function of these medians is then plotted and contrasted with the median value of the specific measure in the original volume.

2.3.4 Results

Inferential properties

Test size Results for the models with correctly and incorrectly specified expected signal are very similar with respect to the test size. In Figure 2.2, the results for the models with correctly specified signal are graphically displayed[2].

Under $AR(1)$ and $AR(2)$, the size of the test for $\beta_1 = 0$ is not different from its nominal level for the fully parametric approach ($\ell = 1$, i.e. whitening procedure), both for a fixed and varying $AR(1)$ coefficient over voxels, and for the parametric approach with additional blocking. For the semi-parametric approach, test sizes are too liberal except for larger block lengths ($\ell > 5$) under $AR(1)$ temporal noise. Under these temporal noise structures, classical parametric (non-bootstrap) inference performs at nominal levels.

Under the complex noise scenario, test sizes for all procedures become too liberal except for the fully parametric approach with varying $AR(1)$ coefficient over voxels which renders conservative results. The test size for classical inference is also not at the nominal level.

We test these patterns more formally via an analysis of variance (ANOVA) model with 3 factors: 1) type of residual (semi-parametric/fully parametric); 2) block length (1-5-20-40 and variable length) and 3) type of noise ($AR(1)$, $AR(2)$ and complex noise). At the $\alpha = 0.05$ significance level we find significant main effects of type of residual $F(1, 8) = 9.73, p < 0.05$

[2]In Table S2.1, in the Supplementary Material, the observed empirical test sizes are presented for both a correctly specified and an incorrectly specified expected signal.

and type of noise $F(2,8) = 24.78, p < 0.001$. The interaction between type of residual and type of noise is also significant ($F(2,8) = 9.54, p < 0.01$). Corrected post-hoc contrasts reveal differences between the residual types in the complex noise condition ($t(8) = 6.948, p < 0.001$).

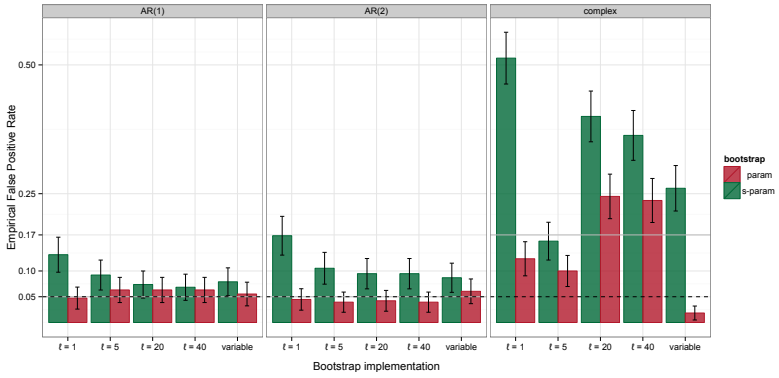


Figure 2.2 Average size and standard error (se) of the test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ in the absence of activation ($\beta_1 = 0$) under a correct specification of the expected signal with $\alpha = 0.05$. The grey line indicates the average size under classical (non-bootstrap) inference. The black dotted line is at α. param: parametric bootstrap; s-param: semi-parametric bootstrap; variabele: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

Power ROC curves that show the empirical true positive rate in function of the empirical false negative rate are displayed in Figure 2.3. To maintain the overview in this figure, we opt to display only the fully ($\ell = 1$) parametric approach with fixed $AR(1)$ coefficient and the (semi-) parametric approaches with $\ell > 5$ next to classical (non-bootstrap) inference, because of their performance with respect to the test size.

For models with correctly specified expected signal, we find similar, good performances of all procedures. Under the complex noise structure, power becomes smaller for all procedures. In that case, we find that the fully parametric bootstrap approach performs best (comparable with the classical inference); there is no advantage of additional blocking.

For models with incorrectly specified mean with $AR(1)$ and $AR(2)$ noise structure, the bootstrap procedures perform better than the classical inference. The highest power is observed for the fully parametric

approach. For the complex noise scenario, poor results are obtained for all procedures.

Similar to the test size, we test these patterns more formally via an ANOVA with Area Under the curve (AUC) as outcome and with 3 factors for the correctly and incorrectly specified HRF model separately: 1) type of residual and 2) block length and 3) noise type. For the correctly specified HRF model we detect significant main effects of type of residual $F(1,8) = 9.73, p < 0.05$ and type of noise $F(2,8) = 24.78, p < 0.001$. The interaction between type of residual and type of noise is also significant ($F(2,8) = 9.54, p < 0.01$). Post-hoc contrasts reveal differences between the noise types in the AR(2) noise condition ($t(8) = 5.659, p < 0.01$).

For the incorrectly specified HRF model we only detect significant main effects of type of residual $F(1,8) = 14.40, p < 0.05$ and type of noise $F(2,8) = 24.78, p < 0.001$. There is no significant interaction between type of residual and type of noise.

Based on these ANOVA models, we construct the 95% Confidence Intervals (CIs) to determine whether the bootstrap procedures differ from classical inference in terms of AUC. If the model for the mean signal is correctly specified, we do not detect differences between the bootstrap approaches and the classical inference in the $AR(1)$ and $AR(2)$ case. For the complex noise, we find the classical approach to be the better approach in all cases, except for the *i.i.d.* semi-parametric bootstrap scenario. When the model of the mean signal is incorrectly specified, for the $AR(1)$ noise, only the 95% CI of the semi-parametric bootstrap with variable block length and $\ell = 40$ contains the AUC of the classical inference, meaning that these are the only cases in which the bootstrap approaches are not better. For the $AR(2)$ noise none of the 95% CI's contain the classical inference value, except for when the variable $AR(1)$ coefficient was used. Again the bootstrap procedures result in a higher AUC. For the complex noise, there are no differences between the the classical inference and the bootstrap approaches. One exception here is when the block length is 5. This holds both for the semi- and the fully parametric approaches, where lower AUCs are observed for the bootstrap procedures.

Image reproducibility

Figure 2.4 provides an overview of all results with respect to image reproducibility. In this figure, the bias of the different bootstrap procedures is compared to the bias of the fully parametric approach with fixed $AR(1)$ coefficient over voxels. This approach is chosen as the reference

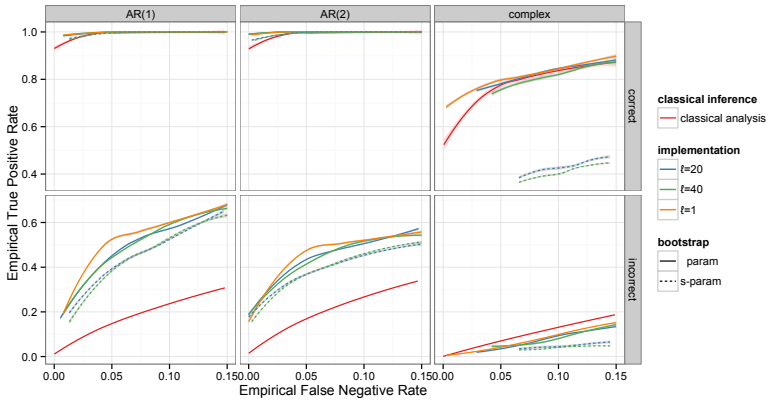
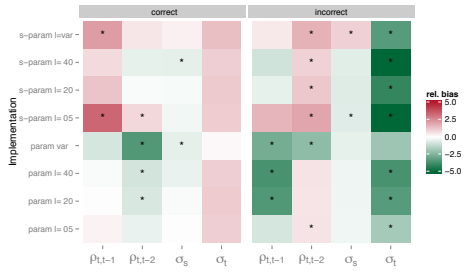


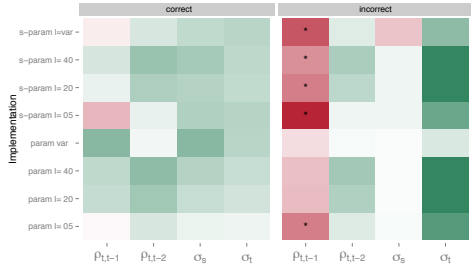
Figure 2.3 Empirical ROC curves for the correct (upper panel) and an incorrect specification of the mean (lower panel). param: parametric bootstrap; s-param: semi-parametric bootstrap

level because it is most the widespread approach to deal with temporally correlated noise (Carp, 2012). The reproducibility of the different image characteristics is discussed in more detail below.

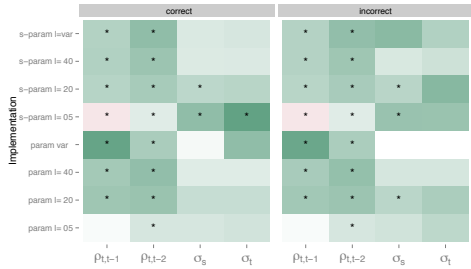
We also test these patterns more formally via an ANOVA model with 3 factors per measure per noise type: 1) type of residual and 2) block length and 3) correctness of the model. Based on this model, we compute a 95% CI around each of the cell means. In Figure 2.4, an asterisk indicates conditions in which the relative bias is significantly different from the relative bias in the fixed AR(1) noise condition.



(a) $AR(1)$ noise.



(b) $AR(1)$ noise.



(c) Complex noise.

Figure 2.4 Relative bias with the fully parametric approach (whitening bootstrap) with fixed $AR(1)$ coefficient as baseline. The relative bias of a bootstrap implementation is calculated as the natural logarithm of the absolute values of the following ratio: (bias bootstrap implementation / bias whitening bootstrap approach with fixed $AR(1)$). Green indicates less bias than such implementation and red indicates more bias. The more intense the color, the more deviation from *whitening* bootstrap with fixed $AR(1)$ coefficient. The asterisk indicates significant differences at the $\alpha = 0.05$ uncorrected level.

$\boldsymbol{\rho_{(t,t-1)}}$ The first order temporal correlation of the residual noise time series is calculated for each voxel (denoted by $\rho_{(t,t-1)}$) and the median is calculated over all voxels in the bootstrapped volumes. Detailed results are presented in Figure 2.5. Again, to maintain the overview in this figure and the following figures, we opt to display only the fully ($\ell = 1$) parametric approach with fixed $AR(1)$ coefficient and the (semi-) parametric approaches with $\ell > 5$ [3].

For models with correct specification of the mean, we find that under $AR(1)$, the parametric bootstrap approach performs best. There is no advantage of varying the $AR(1)$ coefficient over voxels or additional blocking. Under $AR(2)$ and complex noise structure, additional blocking in the parametric approach is needed for a better performance but the best performance is obtained with the fully parametric approach with varying $AR(1)$ coefficient.

For models with incorrect specification of the mean, we find that under $AR(1)$, the parametric approach with additional blocking performs better than the other methods. Under $AR(2)$, the fully parametric approach performs best while the fully parametric approach with varying $AR(1)$ coefficient renders the least biased results under the complex noise structure.

For both correct and incorrect specification of the mean, only under the complex noise structure, the performance of the semi-parametric approach with block length $\ell > 5$ becomes comparable to the parametric approach with additional blocking. There is no further advantage of a variable block length.

$\boldsymbol{\rho_{(t,t-2)}}$ The second order temporal correlation of the residual noise time series is calculated for each voxel (denoted by $\rho_{(t,t-2)}$) and the median is calculated over all voxels in the bootstrapped volumes. Results are presented in Figure 2.6 and additional details are presented in the supplementary material.

For both models with correctly and incorrectly specified means, we find that under $AR(1)$, the fully parametric approach with varying $AR(1)$ coefficient over voxels renders the least biased results. Under $AR(2)$ and complex noise structure, the parametric approach with additional blocking

[3] The graphical depictions of the median values per volume via the smooth densities functions illustrate the performances to reproduce data characteristics for the fully parametric approach and the parametric and semi-parametric approach with block lengths of 20 and 40. Detailed results for all parametric and semi-parametric bootstrap approaches are in the Supplementary Tables S2.2-S2.5

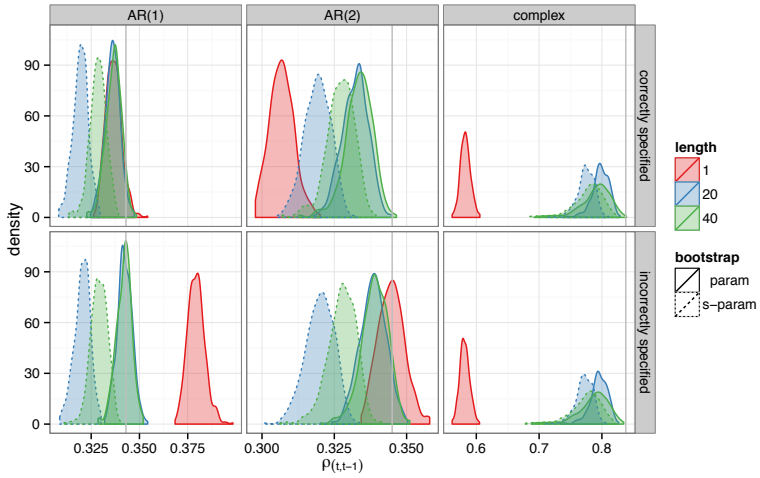


Figure 2.5 Densities of the median $\rho_{(t,t-1)}$ for a correct and incorrect specification of the expected signal. Grey lines indicates the true value. param: parametric bootstrap; s-param: semi-parametric bootstrap

performs better and is similar to the performance of the semi-parametric approach with block length $\ell > 5$. Again, there is no further advantage of a variable block length.

$\boldsymbol{\sigma_t}$ The standard deviation of the residual noise time series is calculated for each voxel (denoted by σ_t) and the median is calculated over all voxels in the bootstrapped volume. Results can be found in Figure 2.7 and additional details are presented in the supplementary material.

For models with correctly specified means, the fully parametric approach performs better than the other procedures under $AR(1)$. Under $AR(2)$, all procedures have similar performance except the fully parametric approach with fixed $AR(1)$ coefficient which renders the most bias. Under the complex noise structure, the semi-parametric approach with small block length ($\ell = 5$) and the fully parametric approach with variable $AR(1)$ coefficient perform best.

For models with incorrectly specified means, the semi-parametric approach and parametric approach with additional blocking perform better than the fully parametric approach.

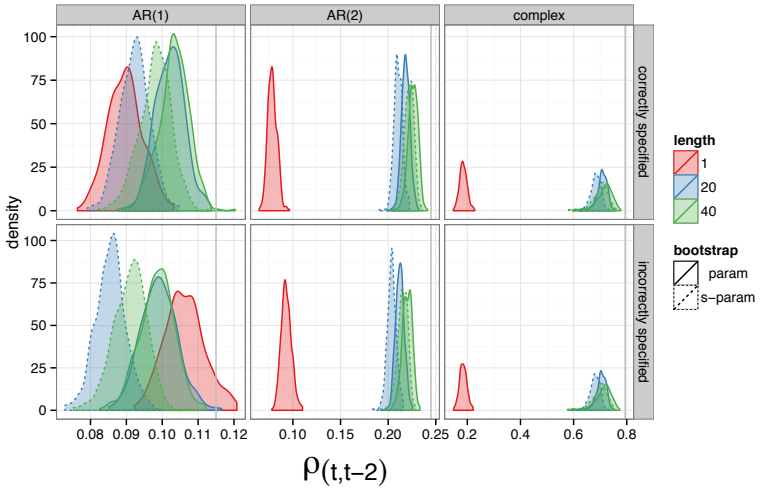


Figure 2.6 Densities of the median $\rho_{(t,t-2)}$ for a correct and incorrect specification of the expected signal. Grey lines indicates the true value. param: parametric bootstrap; s-param: semi-parametric bootstrap

σ_s The standard deviation of the residual noise time series is calculated for each voxel (denoted by σ_s) and the median is calculated over all voxels in the bootstrapped volume. Results are shown in Figure 2.8 and additional details are presented in the supplementary material.

When the mean is correctly specified, under $AR(1)$, we find a similar performance for all procedures. Under $AR(2)$, the fully parametric with varying $AR(1)$ coefficient performs best (compared to no blocking) while the semi-parametric approach with small block length ($\ell = 5$) performs best under the complex noise structure.

The same patterns are observed for models with incorrectly specified means under $AR(1)$ and the complex noise structure. Under $AR(2)$, all procedures have a comparable performance except the semi-parametric approach with variable block length which performs worst.

Summary With respect to inference, the fully parametric approach and the parametric approach with additional blocking perform better than the semi-parametric approach.

With respect to image reproducibility, we find that under a model with correctly specified mean and true underlying $AR(1)$ structure, the fully parametric approach with a fixed $AR(1)$ coefficient performs in general

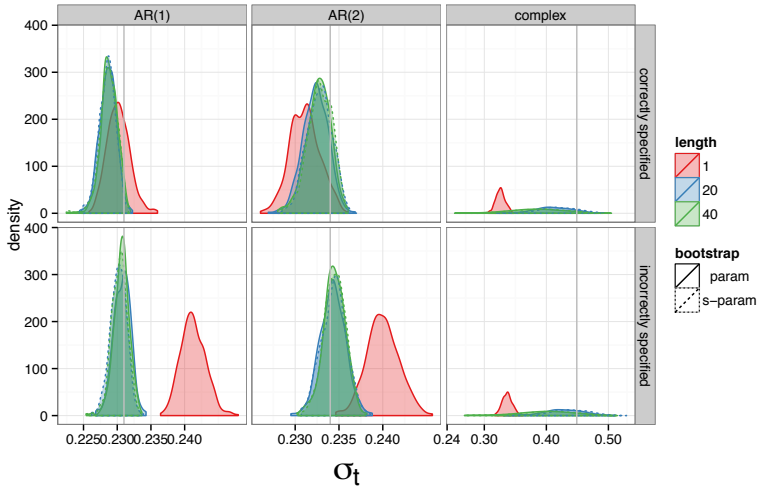


Figure 2.7 Densities of the median σ_t for a correct and an incorrect specification of the expected signal. Grey lines indicates the true value. param: parametric bootstrap; s-param: semi-parametric bootstrap

best as can be expected. For some characteristics, performance becomes even better when allowing a variable $AR(1)$ coefficient over voxels. In almost all other scenarios, we find that the performance of the parametric approach is better with additional blocking with moderate to large block lengths. The apparently flawed performance on $\rho_{t,t-1}$ for the scenario of the misspecified mean is due to the very precise estimation of the fully parametric approach with fixed $AR(1)$.

For image reconstruction, the semi-parametric procedure does not outperform the parametric procedures as in the scenarios with good results, its performance is despite that comparable to the parametric approach with additional blocking.

2.4 Real Data Example

Ishai, Ungerleider, Martin, & Haxby (2000) investigated the representation of objects in the ventral visual pathway by presenting subjects with photographs of faces, houses and chairs. Matching control stimuli with the same visual load were created by scrambling the photographs and line drawings. Two different tasks were performed, passive viewing and

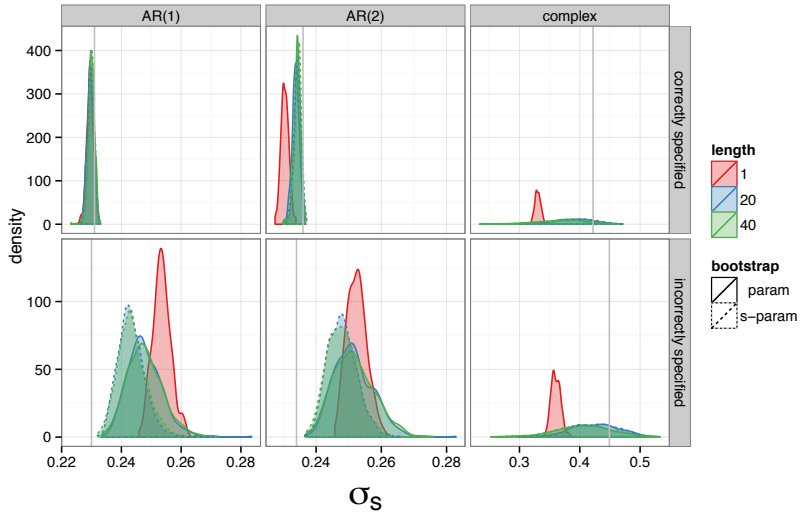


Figure 2.8 Densities of the median σ_s for a correct and an incorrect specification of the expected signal. Grey lines indicates the true value. param: parametric bootstrap; s-param: semi-parametric bootstrap

a delayed match-to-sample task. Here we focus on 4 runs that consisted of 2 times one of the tasks, for one randomly chosen subject (subject 6). Data are in-plane smoothed with a FWHM of 3.75 mm (1 voxel) and the brain extraction proceeded via the FSL6 Brain Extraction Tool (Jenkinson, Pechaud, & Smith, 2005). Linear and quadratic trends are included to account for other temporal artifacts with the fmri package (Tabelow & Polzehl, 2011).

The design matrix consists of the experimental setup and per run we add an intercept to account for baseline level differences. Via a contrast of activation parameters face-related information is compared with the other conditions. We will draw inference using the different bootstrapping procedures discussed in this paper. Also, we add the classical inferential procedure to this comparison. This approach estimates a first order autocorrelation factor per voxel which is then smoothed over the surrounding voxels (Worsley, 2005) with a FWHM of 3.52 mm (default in the fmri package, Tabelow & Polzehl, 2011). These estimates were also used for the fully parametric approach.

Figure 2.9 shows the results that are not corrected for multiple testing, two main findings emerge. First, with increasing block length, the

semi-parametric approach results in less activation. Secondly, the fully parametric approach reveals more activation than the semi-parametric. However, given the necessity to account for the multiplicity in an fMRI test setting, we also demonstrate two possible approaches to deal with multiple testing, i.e. a Family-Wise Error (FWE) correction based on the Westfall-Young procedure (Westfall & Young, 1993; Adolf et al., 2014) and a correction based on the rationale of Lieberman & Cunningham (2009).

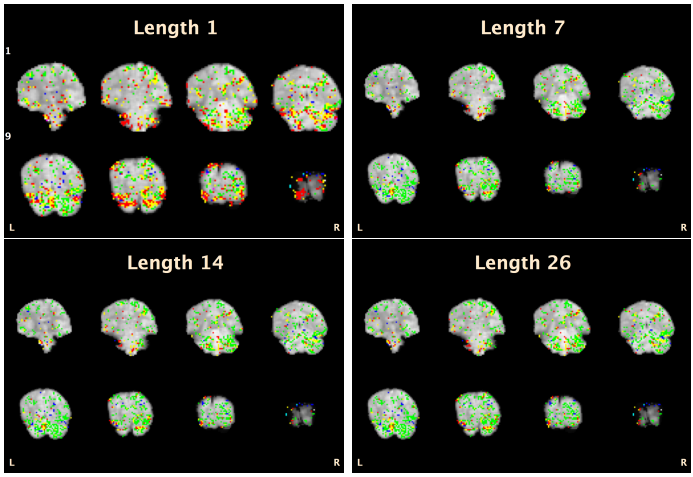


Figure 2.9 Illustration of uncorrected bootstrap-based inference for the data of Ishai et al. (2000), plotted using Mango software (University of Texas Health Science Center). Note that only slices $z = 1-3-5-7-9-11-13-15$ are shown. Yellow: declared active by both the semi- and fully parametric bootstrap; Red: declared active by the semi-parametric bootstrap only; Green: declared active by the fully parametric bootstrap only; Blue: declared active by the classical inference procedure; Purple: declared active by both the classical inference and the semi-parametric approach;and Turquoise: declared active by both the classical inference and the fully parametric bootstrap.

To apply the FWE correction, we construct a null distribution of the maximum t statistic over all voxels. To this end, we bootstrap from the residuals (e or $\tilde{e}$) with no activation. In each bootstrap sample the maximum test statistic is then computed. This procedure is repeated for 5.000 bootstrap samples. The p-values are then computed as in Adolf et al. (2014). The two step procedure of Lieberman & Cunningham (2009) on the other hand aims at a better balance between type I and type II errors

by first applying an uncorrected threshold (e.g. $\alpha = 0.005$). In a second step, it uses a cluster forming algorithm on the above threshold voxels and defines a minimum cluster size (e.g. 10). We employ this method based on 10.000 bootstrap samples to determine the activation via the confidence interval described in the simulation study. We then show the three largest clusters as an illustration of the principle.

Based on the FWE corrected results we did not find any activation at $\alpha = 0.05$ or $\alpha = 0.1$. However, further inspection of the obtained p-values reveals that with increasing block length the semi-parametric approach becomes less anti-conservative compared to the fully parametric approach (see Figure 2.10).

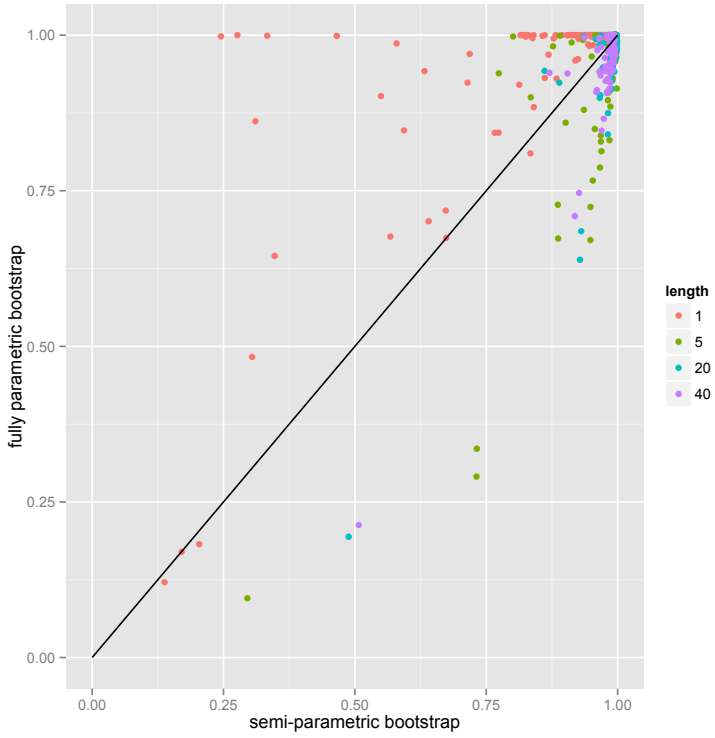


Figure 2.10 FWE-corrected $p$values based on the semi- and fully parametric bootstrap.

In Figure 2.11, we compare the activation patterns using the second

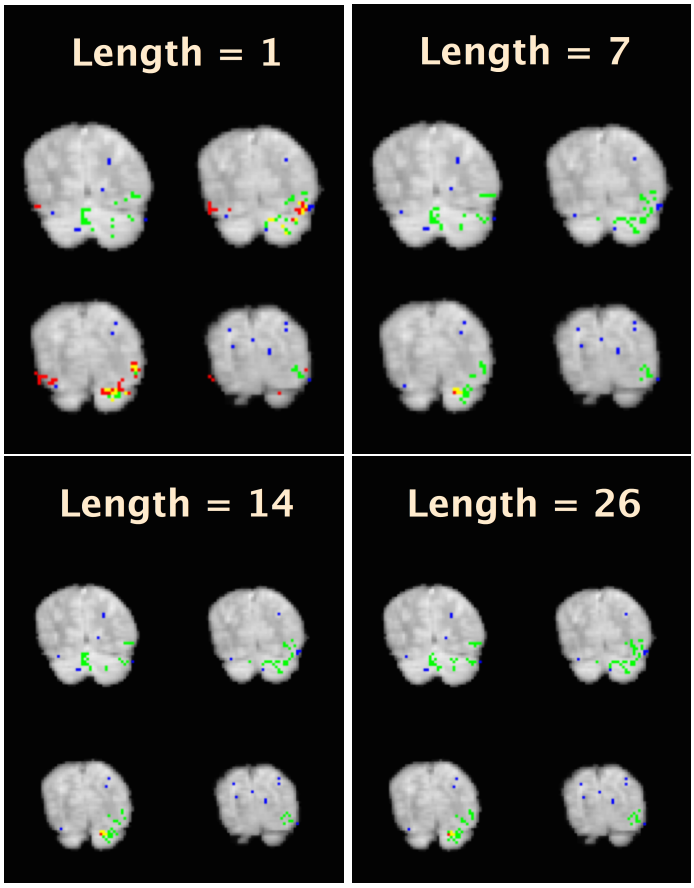


Figure 2.11 Illustration of bootstrap-based inference for the data of Ishai et al. (2000), plotted using Mango software (University of Texas Health Science Center) using the procedure of Lieberman & Cunningham (2009). Note that only slices $z = 9-12$ are shown. Yellow: declared active by both the semi- and fully parametric bootstrap; Red: declared active by the semi-parametric bootstrap only; Green: declared active by the fully parametric bootstrap only; Blue: declared active by the classical inference procedure.

approach. For illustrational purposes only z-slices 9 to 12 are depicted. The extent of the network differs between the parametric and the semi-parametric bootstrap approaches. It also differs with increasing ℓ. For the semi-parametric approach, there is a tendency for declaring a decreasing number of voxels as active with increasing block lengths. For the fully-

parametric approach, there is a slight tendency for declaring an increasing number of voxels as active with increasing block lengths. From Figure 2.11 it is also clear that the fully parametric approach results in more activation.

Next, we assess the reproducibility properties of the bootstrap procedures. We focus on the pattern of temporal correlation within the bootstrap samples in comparison with the original data. We consider parametric and semi-parametric bootstrap approaches with block lengths of $1, 7, 14$ and 26. For the determination of the $AR(1)$ coefficient in the fully parametric approach, the implementation of Worsley (2005) was used with the above described smoothing procedure. For the determination of $\ell_{opt,i}$ only very small block lengths were found making it practically difficult to obtain good recovery of the temporal properties. We did not include this scenario in our comparison.

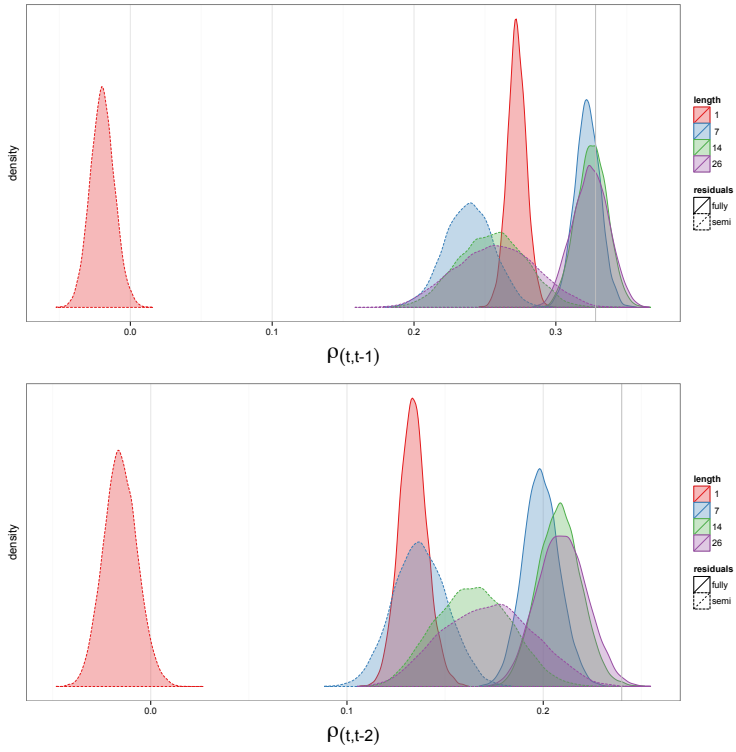


Figure 2.12 The median $\rho_{t,t-1}$(upper panel) $\rho_{t,t-2}$ (lower panel) and in the real data.

Figure 2.12 shows the median of the estimated first and second order temporal correlation per bootstrap sample and contrasts it with the observed median first order temporal correlation (which is based on an assumed correct specification of the mean structure). Both the semi-parametric and the fully parametric bootstrap retain the temporal correlation well, if a sufficient amount of blocking is used. We find a slight underestimation of the temporal correlation used by the fully parametric bootstrap. This is however resolved with additional blocking ($\ell > 5$). The under-estimation is larger for the second order correlation than for the first order correlation.

2.5 Discussion

There is emerging evidence that a full parameterization in GLM-based data analysis for fMRI may not lead to valid inference (Eklund et al., 2012; Lenoski et al., 2008). In the current study a simulation-based and an empirical comparison of two GLM-based methods is conducted to further unravel the abilities of bootstrap procedures.

Concerning inferential properties we demonstrate that if the parameterization of the noise model (almost) holds, it is safe to rely on a full parameterization. However, given an appropriate block length, we demonstrate that under simple noise structures parametric and semi-parametric bootstrap procedures are valuable alternatives. A test size equal to the nominal level for the parametric approach is achieved when the mean signal is correctly specified. This is not the case for the semi-parametric approach where we observe anti-conservative test sizes. For complexly structured noise, we advise cautiousness since inferential conclusions are demonstrated to be precarious. This finding is similar to Tang et al. (2008), who showed that test sizes based on wavelet bootstrapping are too liberal under complex temporal noise.

We have also explored the effects of a misspecified mean signal. In that case the performance of classical parametric (non-bootstrap) inference steadily dropped. By contrast, we show that GLM-based bootstrap approaches (both parametric and semi-parametric) allow for better informed inferential decisions. This further puts evidence to bootstrapping as a proper way of conducting inference in fMRI studies (see e.g. Darki & Oghabian, 2013). We stress however that inference based on bootstrap is in the case of complex noise structures not at nominal levels regardless of correct specification of the mean signal.

In the real data example we find moreover more activation using the procedure of Lieberman & Cunningham (2009) with bootstrapping than with classical parametric testing using the t distribution. This confirms the results of the AUC analysis. In addition, we also demonstrate the elegant and broad applicability of the bootstrap principle for multiple testing via FWE corrected p-values. Adolf et al. (2014) used a highly similar approach called block-wise permutation. While permutation methods allow for *exact* (corrected) inference (if all permutations are determined and full exchangeability is guaranteed), bootstrap procedures serve the same purpose, albeit approximate in nature (Nichols & Hayasaka, 2003).

Next to inferential properties, we also explored reproducibility properties of the bootstrap. In general, images are well reproduced (i.e. they mimic the original image) with a fully parametric bootstrap approach. Its performance can however be substantially improved by additional blocking when the noise and/or the expected signal is misspecified. Although a semi-parametric approach cannot outperform such parametric approach with additional blocking, it assures good reproducibility. The semi-parametric procedure offers a clear computational advantage. One semi-parametric bootstrap cycle took on average $8.05s$ to complete the analysis of a voxel compared to $34.67s$ per cycle for the fully parametric bootstrap. The classical analysis, relying on the parametric t-distribution takes roughly as long as the fully parametric bootstrap. For 10.000 bootstrap samples this difference in computation time between the fully parametric and the semi-parametric bootstrap procedure can amount to more than 72 hours on a single-core processor.

Using a semi-parametric approach or a fully parametric approach with additional blocking imposes the end-user to define a block length ℓ. An appropriate choice of such block length is thus a key determinant in applying these approaches. Indeed, under the simple noise scenarios the performance of the semi-parametric approach clearly improved with increasing block length. It is therefore strongly recommended to reduce the noise in the GLM as much as possible in order to decrease the noise complexity. This could for example be achieved by taking into account for example physiological noise sources (Welvaert & Rosseel, 2012).

Finally, we note that a variable block length (which varies *over* voxels) does not necessarily improve the ability to reproduce data characteristics of the original data set. We suggest an easy ad-hoc rule of thumb for the selection of ℓ: take from all estimates $\ell_{opt,i}$ the 85% quantile of that distribution over the voxels. Using this rule, the block lengths were determined

7,11 and 35 for the $AR(1)$, $AR(2)$ and complex noise which is close to the block lengths considered here. These generally resulted in a good overall performance in the simulations.

2.6 Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University, the Hercules Foundation and the Flemish Government department EWI.

Supplementary Material

Table S2.1 Average size and standard error (se) of the test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ in the absence of activation ($\beta_1 = 0$) under a correct and an incorrect specification of the mean signal with $\alpha = 0.05$. param: parametric bootstrap; s-param: semi-parametric bootstrap; var: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

		$\ell = 1$		$\ell = 2$		$\ell = 5$		$\ell = 20$		$\ell = 40$		var	
		size	se	size	se	size	se	size	se	size	se	size	se
correct													
$AR(1)$	s-param	0.132	0.017	0.103	0.016	0.092	0.015	0.074	0.013	0.068	0.013	0.079	0.014
	param	0.047	0.011	0.055	0.012	0.063	0.012	0.063	0.012	0.063	0.012	0.055	0.012
$AR(2)$	s-param	0.168	0.019	0.132	0.017	0.105	0.016	0.095	0.015	0.095	0.015	0.087	0.014
	param	0.045	0.011	0.061	0.012	0.039	0.010	0.042	0.010	0.039	0.010	0.061	0.012
complex	s-param	0.513	0.026	0.345	0.024	0.158	0.019	0.400	0.025	0.363	0.025	0.261	0.023
	param	0.124	0.017	0.118	0.017	0.100	0.015	0.245	0.022	0.237	0.022	0.018	0.007
incorrect													
$AR(1)$	s-param	0.132	0.017	0.103	0.016	0.092	0.015	0.074	0.013	0.068	0.013	0.089	0.015
	param	0.039	0.010	0.047	0.011	0.058	0.012	0.061	0.012	0.058	0.012	0.055	0.012
$AR(2)$	s-param	0.168	0.019	0.124	0.017	0.084	0.014	0.074	0.013	0.071	0.013	0.082	0.014
	param	0.042	0.010	0.050	0.011	0.058	0.012	0.058	0.012	0.061	0.012	0.047	0.011
complex	s-param	0.513	0.026	0.345	0.024	0.158	0.019	0.400	0.025	0.363	0.025	0.287	0.023
	param	0.124	0.017	0.116	0.016	0.100	0.015	0.245	0.022	0.239	0.022	0.021	0.007

Table S2.2 Median $\rho_{(t,t-1)}$ and standard error (se) for a correct and incorrect specification of the mean signal. param: parametric bootstrap; s-param: semi-parametric bootstrap; var: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

		true	$\ell = 1$		$\ell = 2$		$\ell = 5$		$\ell = 20$		$\ell = 40$		var	
	residuals	true	med	se	med	se	med	se	med	se	med	se	med	se
correct														
$AR(1)$	s-param	0.343	−0.005	0.005	0.166	0.004	0.268	0.004	0.320	0.004	0.329	0.004	0.288	0.004
	param	0.343	0.336	0.004	0.334	0.004	0.335	0.003	0.336	0.004	0.337	0.004	0.340	0.004
$AR(2)$	s-param	0.345	−0.005	0.005	0.167	0.004	0.267	0.004	0.319	0.005	0.328	0.005	0.289	0.004
	param	0.345	0.307	0.004	0.300	0.004	0.320	0.004	0.333	0.004	0.334	0.005	0.341	0.004
complex	s-param	0.838	−0.006	0.022	0.412	0.015	0.661	0.011	0.776	0.014	0.782	0.024	0.780	0.003
	param	0.838	0.583	0.008	0.614	0.011	0.742	0.009	0.798	0.014	0.794	0.025	0.825	0.003
incorrect														
$AR(1)$	s-param	0.343	−0.005	0.005	0.166	0.005	0.268	0.004	0.321	0.004	0.329	0.004	0.285	0.004
	param	0.343	0.379	0.004	0.359	0.004	0.348	0.003	0.342	0.004	0.342	0.004	0.341	0.005
$AR(2)$	s-param	0.345	−0.005	0.005	0.167	0.005	0.269	0.004	0.320	0.005	0.328	0.005	0.286	0.004
	param	0.345	0.345	0.005	0.320	0.004	0.331	0.004	0.339	0.004	0.339	0.005	0.342	0.006
complex	s-param	0.837	−0.007	0.021	0.410	0.015	0.659	0.011	0.773	0.014	0.778	0.025	0.778	0.003
	param	0.837	0.581	0.008	0.612	0.011	0.740	0.009	0.796	0.014	0.790	0.026	0.823	0.003

Table S2.3 Median $\rho_{(t,t-2)}$ and standard error (se) for an correct and incorrect specification of the mean signal. param: parametric bootstrap; s-param: semi-parametric bootstrap; var: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

		true	$\ell = 1$		$\ell = 2$		$\ell = 5$		$\ell = 20$		$\ell = 40$		var	
	residuals	true	med	se	med	se	med	se	med	se	med	se	med	se
correct														
$AR(1)$	s-param	0.113	−0.004	0.005	−0.005	0.006	0.059	0.004	0.092	0.004	0.098	0.004	0.074	0.004
	param	0.113	0.090	0.005	0.089	0.005	0.097	0.005	0.102	0.004	0.103	0.004	0.114	0.005
$AR(2)$	s-param	0.245	−0.003	0.005	−0.005	0.006	0.138	0.005	0.210	0.004	0.222	0.005	0.169	0.004
	param	0.245	0.078	0.005	0.076	0.006	0.168	0.005	0.218	0.004	0.226	0.005	0.114	0.005
complex	s-param	0.794	−0.004	0.022	−0.008	0.030	0.462	0.021	0.685	0.019	0.706	0.031	0.723	0.004
	param	0.794	0.184	0.014	0.209	0.022	0.522	0.019	0.709	0.018	0.721	0.032	0.679	0.005
incorrect														
$AR(1)$	s-param	0.115	−0.004	0.005	−0.005	0.006	0.054	0.005	0.086	0.004	0.091	0.005	0.068	0.004
	param	0.115	0.106	0.006	0.099	0.006	0.098	0.005	0.099	0.005	0.099	0.005	0.114	0.006
$AR(2)$	s-param	0.245	−0.004	0.005	−0.005	0.006	0.133	0.005	0.204	0.005	0.215	0.005	0.163	0.004
	param	0.245	0.093	0.005	0.084	0.006	0.166	0.005	0.212	0.005	0.220	0.005	0.115	0.006
complex	s-param	0.794	−0.003	0.021	−0.008	0.029	0.460	0.020	0.681	0.019	0.701	0.032	0.721	0.004
	param	0.794	0.184	0.014	0.208	0.021	0.519	0.019	0.706	0.019	0.717	0.033	0.676	0.005

Table S2.4 Median σ_t and standard error (se) for a correct and incorrect specification of the mean signal. param: parametric bootstrap; s-param: semi-parametric bootstrap; var: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

		true	$\ell = 1$		$\ell = 2$		$\ell = 5$		$\ell = 20$		$\ell = 40$		*var*	
	residuals	true	med	se	med	se	med	se	med	se	med	se	med	se
correct														
$AR(1)$	s-param	0.231	0.229	0.001	0.229	0.001	0.229	0.001	0.229	0.001	0.229	0.001	0.228	0.001
	param	0.231	0.230	0.002	0.229	0.002	0.229	0.001	0.229	0.001	0.229	0.001	0.230	0.002
$AR(2)$	s-param	0.234	0.234	0.001	0.234	0.001	0.233	0.001	0.233	0.001	0.233	0.001	0.233	0.001
	param	0.234	0.231	0.002	0.229	0.002	0.231	0.001	0.232	0.001	0.233	0.001	0.235	0.002
complex	s-param	0.451	0.454	0.009	0.451	0.011	0.445	0.017	0.420	0.031	0.389	0.045	0.395	0.009
	param	0.451	0.326	0.008	0.348	0.009	0.401	0.015	0.410	0.031	0.384	0.044	0.437	0.010
incorrect														
$AR(1)$	s-param	0.230	0.231	0.001	0.231	0.001	0.230	0.001	0.230	0.001	0.230	0.001	0.230	0.001
	param	0.230	0.241	0.002	0.235	0.002	0.232	0.002	0.231	0.001	0.231	0.001	0.232	0.002
$AR(2)$	s-param	0.234	0.236	0.001	0.235	0.001	0.235	0.001	0.234	0.001	0.234	0.001	0.235	0.001
	param	0.234	0.240	0.002	0.234	0.002	0.234	0.001	0.234	0.001	0.234	0.001	0.237	0.002
complex	s-param	0.449	0.474	0.010	0.471	0.012	0.465	0.018	0.438	0.032	0.407	0.045	0.474	0.010
	param	0.449	0.337	0.008	0.360	0.010	0.418	0.015	0.427	0.031	0.400	0.045	0.337	0.008

Table S2.5 Median σ_s and standard error (se) for an correct and incorrect specification of the mean signal. param: parametric bootstrap; s-param: semi-parametric bootstrap; var: ℓ_{opt} for semi-parametric bootstrap and variable $AR(1)$ for fully parametric bootstrap.

		true	$\ell = 1$		$\ell = 2$		$\ell = 5$		$\ell = 20$		$\ell = 40$		var	
	residuals	true	med	se	med	se	med	se	med	se	med	se	med	se
correct														
$AR(1)$	s-param	0.231	0.230	0.001	0.230	0.001	0.230	0.001	0.230	0.001	0.230	0.001	0.229	0.001
	param	0.231	0.230	0.001	0.229	0.001	0.229	0.001	0.230	0.001	0.230	0.001	0.230	0.001
$AR(2)$	s-param	0.236	0.235	0.001	0.235	0.001	0.234	0.001	0.234	0.001	0.235	0.001	0.234	0.001
	param	0.236	0.230	0.001	0.229	0.001	0.232	0.001	0.234	0.001	0.234	0.001	0.236	0.001
complex	s-param	0.422	0.418	0.009	0.415	0.012	0.412	0.016	0.397	0.031	0.376	0.041	0.467	0.007
	param	0.422	0.330	0.005	0.336	0.008	0.381	0.014	0.392	0.031	0.372	0.042	0.500	0.007
incorrect														
$AR(1)$	s-param	0.230	0.243	0.003	0.243	0.003	0.243	0.004	0.243	0.005	0.243	0.005	0.292	0.003
	param	0.230	0.253	0.003	0.249	0.004	0.248	0.005	0.247	0.006	0.247	0.006	0.245	0.003
$AR(2)$	s-param	0.234	0.248	0.003	0.248	0.003	0.248	0.004	0.248	0.004	0.248	0.005	0.298	0.004
	param	0.234	0.252	0.003	0.247	0.004	0.250	0.005	0.251	0.006	0.251	0.006	0.251	0.003
complex	s-param	0.449	0.440	0.012	0.438	0.015	0.437	0.021	0.426	0.038	0.407	0.046	0.440	0.012
	param	0.449	0.359	0.008	0.366	0.011	0.413	0.019	0.426	0.040	0.409	0.048	0.359	0.008

References

Adolf, D., Weston, S., Baecke, S., Luchtmann, M., Bernarding, J., & Kropf, S. (2014). Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in neuroinformatics*, *8*, 72.

Aguirre, G. K., Zarahn, E., & D'esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *NeuroImage*, *8*(4), 360–9.

Bellec, P., Perlbarg, V., & Evans, A. C. (2009). Bootstrap generation and evaluation of an fMRI simulation database. *Magnetic resonance imaging*, *27*(10), 1382–96.

Bullmore, E., Fadili, J., Maxim, V., Sendur, L., Whitcher, B., Suckling, J., ... Breakspear, M. (2004). Wavelets and functional magnetic resonance imaging of the human brain. *NeuroImage*, *23 Suppl 1*, S234–49.

Buxton, R. B., Uludağ, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, *23 Suppl 1*, S220–33.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Chatfield, C. (2000). *The analysis of time series. an introduction.* (C. Chatfield & J. V. Zidek, Eds.). Boca Raton: Chapman & Hall/CRC.

Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61.

Darki, F., & Oghabian, M. A. (2013). False positive control of activated voxels in single fMRI analysis using bootstrap resampling in comparison to spatial smoothing. *Magnetic resonance imaging*, *31*(8), 1331–1337.

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their application.* Cambridge: University Press.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Friman, O., & Westin, F.-J. (2005). Resampling fmri time series. *NeuroImage*, *25*, 859-867.

Glaser, D., & Friston, K. J. (2007). Covariance components. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 140-147). Elsevier Ltd./Academic Press.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, *9*(4), 416–29.

Greve, D., Brown, G., Mueller, B., Glover, G., & Liu, T. (2012). A survey of the sources of noise fMRI. *Psychometrika*, doi: 10.1007/S11336-012-9294-0.

Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, *43*(3), 509–20.

Gudbjartsson, H., & Patz, S. (2005). The Rician Distribution of Noisy MRI Data. *Magnetic Resonance in Medicine*, *34*(6), 910–914.

Hayfield, T., & Racine, J. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, *27*, 5: 1-32.

Henson, R., & Friston, K. J. (2007). Convolution models for fmri. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 193-210). Elsevier Ltd./Academic Press.

Ishai, A., Ungerleider, L. G., Martin, a., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of cognitive neuroscience*, *12 (S2)*, 35–51.

Jenkinson, M., Pechaud, M., & Smith, S. M. S. (2005). Bet2: Mr-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping.*

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear models.* New York: McGraw-Hill Irwin.

Lahiri, S. N. (2003). *Resampling methods for dependent data.* Springer-Verlag, Inc.

Laird, A. R., Rogers, B. P., & Meyerand, M. E. (2004). Comparison of Fourier and wavelet resampling methods. *Magnetic Resonance in Medicine*, *51*(2), 418–22.

Lazar, N. A. (2008). Noise and data preprocessing. In *The statistical analysis of functional mri data* (p. 37-51). Springer.

Lenoski, B., Baxter, L. C., Karam, L. J., Maisog, J., & Debbins, J. (2008). On the performance of autocorrelation estimation algorithms for fmri analysis. *IEEE Journal of Selected Topics in Signal Processing*, *2*, 828-838.

Lieberman, M. D., & Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, *4*(4), 423–8.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464.

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mismodeling. *NeuroImage*, *45*(1 Suppl), S187–98.

Luo, W.-L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, *19*, 1014-1032.

Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in human neuroscience*, *5*, 28.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Politis, D. N., & White, H. (2004). Automatic Block-Length Selection for the Dependent Bootstrap. *Econometric Reviews*, *23*(1), 53–70.

Razavi, M., Grabowski, T. J., Vispoel, W. P., Monahan, P., Mehta, S., Eaton, B., & Bolinger, L. (2003). Model assessment and model building in fMRI. *Human brain mapping*, *20*(4), 227–38.

R Core Team, . (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Tabelow, K., & Polzehl, J. (2011). Statistical Parametric Maps for Functional MRI Experiments in R : The Package fmri. *Journal of Statistical Software*, *44*(11).

Tang, L., Woodward, W. a., & Schucany, W. R. (2008). Undercoverage of Wavelet-Based Resampling Confidence Intervals. *Communications in Statistics - Simulation and Computation*, *37*(7), 1307–1315.

Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstädt-Klein, S., ... Poline, J.-B. (2012). Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. *NeuroImage*, *61*(1), 295–303.

Wellcome Trust Centre for Neuroimaging U.C.L. (2010). Spm 8 [Computer software manual]. http:// www.fil.ion.ucl.ac.uk/spm/.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaeghe, G., & Rosseel, Y. (2011). Journal of Statistical Software. *Journal of Statistical Software*, *44*(10), 10: 1-18.

Welvaert, M., & Rosseel, Y. (2012). How ignoring physiological noise can bias the conclusions from fMRI simulation results. *Journal of neuroscience methods*, *211*(1), 125–32.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing. examples and methods for p-value adjustment.* John Wiley Sons, Inc.

Worsley, K. J. (2005). Spatial smoothing of autocorrelations to control the degrees of freedom in fMRI analysis. *NeuroImage*, *26*(2), 635–41.

Zhang, H., Luo, W.-L., & Nichols, T. E. (2006). Diagnosis of single-subject and group fMRI data with SPMd. *Human brain mapping*, *27*(5), 442–51.

3 Data Analytical Stability of Cluster-wise and Peak-wise Inference in fMRI Data Analysis

Abstract Carp (2012) demonstrated the large variability that is present in the method sections of fMRI studies. This methodological variability between studies limits reproducible research.

Evaluation protocols for methods used in fMRI should include data analytical stability measures quantifying the variability in results following choices in the methods. Data analytical stability can be seen as a proxy for reproducibility. To illustrate how one can perform such evaluations, we study two competing approaches for topological feature based inference (random field theory and permutation-based testing) and two competing methods for smoothing (Gaussian smoothing and adaptive smoothing). We compare these approaches from the perspective of data analytical stability in real data, and additionally consider validity and reliability in simulations. There is clear evidence that choices in the methods impact the validity, reliability and stability of the results. For the particular comparison studied, we find that permutation-based methods render the most valid results. For stability and reliability, the performance of different smoothing and inference types depends on the setting. However, while being more reliable, adaptive smoothing can evoke less stable results when using larger kernel width, especially with cluster size based permutation inference. While existing evaluation methods focus on validity and reliability, we show that data analytical stability enables to further distinguish between performance of different methods. Data analytical stability is an important additional criterion that can easily be incorporated in evaluation protocols.

This chapter has been published in The Journal of Neuroscience Methods.

Roels, S. P., Bossier, H., Loeys, T., & Moerkerke, B. (2015). Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis. *Journal of Neuroscience Methods, 240*, 37–47.

3.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) has emerged over the years as a prominent tool to investigate and localize brain functions. To select brain regions that are activated during specific tasks, the brain is divided in artificially created cubicles or voxels. For these voxels, the Blood Oxygen Level Dependent (BOLD) signal is measured on a series of time points. The BOLD signals during task processing are contrasted with signals during rest or during the performance of another task. In the commonly used General Linear Model (GLM) approach, a linear model is fitted for each voxel after which the results are visualized in a three dimensional Statistical Parametric Map (SPM). This map shows the test statistic for each voxel (Friston et al., 1995; Worsley et al., 2002; Ashby, 2011), allowing to evaluate evidence for activation.

The selection of active data features (clusters or peaks) is not limited to drawing inference about activation but can be viewed as a sequence of four phases (see e.g. Friston, Ashburner, Kiebel, Nichols, & Penny, 2007; Lindquist, 2008; Worsley et al., 2002). In the first phase, the study design is set up, and one must ensure the estimability of the effects of interests (Smith, Jenkinson, Beckmann, Miller, & Woolrich, 2007). Next, noise-reducing pre-processing steps take place (see e.g. Friston et al., 2007). In the third phase, the modeling is typically conducted via a GLM procedure (Friston et al., 1995), although data-driven approaches such as Independent Compontent Analysis (ICA) and Principle Component Analysis (PCA) are rapidly gaining popularity too over the last couple of years (e.g. Beckmann, 2012; Viviani, Grön, & Spitzer, 2005). In the final phase, the inferential phase, the evidence for activation is sought (Worsley, Taylor, Tomaiuolo, & Lerch, 2004; Nichols & Holmes, 2002). This sequence of four phases will further be referred to as the *selection procedure*.

Carp (2012) recently demonstrated the large variation in such selection procedures. Each of these four phases is known to have an influence on the selected features and thus impacts the comparability of results (see e.g. Bennett & Miller, 2013; Della-Maggiore, Chau, Peres-Neto, & McIntosh, 2002; Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012; Nichols & Hayasaka, 2003, for effects on respectively the design phase, the pre-processing phase, the modeling phase and the inferential phase). However, the effect of each of the choices on the reproducibility of the selected features is rarely taken into account when evaluating selection procedures. One notable exception is the Nonparametric, Prediction, Activation, In-

fluence, Reproducibility, re-Sampling (NPAIRS) framework (e.g. Strother et al., 2002, 2004) which aims at optimizing data analytic pipelines. In a cross-validation protocol this framework quantifies reproducibility via the reliability among subsampled images.

With Bennett & Miller (2013), who advocated the imperative search for reproducible methods in the context of fMRI methods, we argue here it is important to investigate the *stability* of different choices in the selection procedure too. The concept of stability was first introduced in the context of selecting genes associated with a phenotype (Qiu, Xiao, Gordon, & Yakovlev, 2006). These authors quantified stability as a selection criterion through the variability on the number of selected genes and the frequency with which these genes are detected in different samples. The composition of the candidate genes list is subject to random fluctuations and thus finding differently expressed genes is subject to random fluctuations as well. The higher the variability on the number of selected genes, the smaller the stability. Similarly, genes that are only selected to be associated with a phenotype in a limited number of samples are indicators for a smaller stability. The concept of stability can be translated to the context of fMRI. Due to the selection procedure in fMRI, the set of candidate features is subject to random fluctuations too, resulting in variability. Stability thus refers to the ability to replicate the selected data features in a replication context. This variability is often ignored when evaluating methods in fMRI research (see however Durnez, Roels, & Moerkerke, 2014; Strother et al., 2002).

Indeed, traditional evaluation protocols merely involve an assessment of *validity* by verifying whether the type I error rate is controlled at the nominal level (e.g. Hayasaka & Nichols, 2003). This can be done by inspection of the type I error rates and the distribution of the p-values under H_0. Recently, larger focus has been put on the type II error rate (Button et al., 2013) and *reliability* (e.g. Wilke, 2012; Gorgolewski, Storkey, Bastin, & Pernet, 2012). The latter can for example be investigated by inspecting how close the selected features are to the truely activated brain regions. In this paper we argue that a third aspect, the *data analytical stability*, has to be equally valued during the evaluation of methods for the selection of brain activation. This stability can be verified by an inspection of the variation in the selected features over repeated samples. All three aspects strongly impact the ability to obtain reproducible results.

We thus focus in this paper on the exploration of reproducibility in terms of validity, reliability and stability in the context of single-subject

fMRI. While the first two concepts have already been extensively discussed in the neuroimaging literature, we will show how the latter can be assessed both in simulation context as on real data. To illustrate the concept, we will focus on particular choices made in two of the aforementioned four phases: the pre-processing phase and the inferential phase. We anticipated the largest impact of choices in these 2 phases but the presented approach can easily accommodate choices in the other two phases too. In this paper, we opted to fix the design phase and chose the popular GLM-approach for the modeling phase. From an inferential point of view, it is important to note that simultaneously testing for activation in each voxel induces a large multiple testing problem. One way to circumvent this problem is to proceed with testing for activation in topological features such as local maxima, called peaks, or spatially neighbouring voxels, called clusters (Hayasaka, Phan, Liberzon, Worsley, & Nichols, 2004; Worsley et al., 2004). This decreases the dimension of the test problem substantially. To this end, two methods are frequently used. Following a theory based set of assumptions, Random Field Theory allows to construct a distribution for peaks and clusters under the null hypothesis of no activation. Non-parametric permutation-based methods on the other hand empirically construct this distribution conditional on the observed data. Since Hayasaka & Nichols (2003) pointed out that the assumptions and robustness of these methods have not been extensively validated, further investigation seems apposite. Since appropriate smoothing is an important prerequisite for applying RFT, we will focus on that aspect in the pre-processing phase. More precisely, we will thus evaluate two approaches to 1) data smoothing (pre-processing phase): classical isotropic Gaussian smoothing (Friston et al., 2007), and a more data-driven adaptive smoothing procedure (Tabelow, Polzehl, Voss, & Spokoiny, 2006); and 2) inference (inferential phase): relying on parametric Random Field Theory (RFT) (e.g. Worsley et al., 1996; Friston et al., 2007) and relying on the empirical properties of the permutation (PERM) null distribution (Nichols & Holmes, 2002). In the next section, we describe these different steps in the selection procedure in more detail and explain the specific implementation of the evaluation protocol in terms of validity, reliability and stability in a simulation study where the underlying truth is known. In section 3, we present the results from this simulation study. Next we show how the evaluation of stability can be assessed in a real data example and link results to findings from the simulation setting. We end with a discussion.

3.2 Choices in the selection procedure and assessment of reproducibility

3.2.1 Topological Inference

In the commonly used mass-univariate approach to single-subject fMRI, a GLM is fitted for the BOLD signal of each voxel over time $\mathbf{Y_v}$ (with $\mathbf{Y_v} : Y_{v1}, \ldots Y_{vt}, \ldots Y_{vT}$, T number of time points, for voxel v with $v = 1, \ldots, V$ the total number of voxels in the brain volume), i.e. (see e.g. Kiebel & Holmes, 2007; Poline & Brett, 2012),

$$\mathbf{Y_v} = \mathbf{X}\boldsymbol{\beta_v} + \boldsymbol{\epsilon_v}, \tag{3.1}$$

$\mathbf{X}$ is the matrix that represents the expected BOLD signal under brain activation, i.e. a convolution of the stimulus onset function with a hemodynamic response function (HRF) (Henson & Friston, 2007). $\boldsymbol{\epsilon_v}$ is the vector representing the residuals per voxel v. The parameters of interest, in vector $\boldsymbol{\beta_v}$, are typically estimated for each voxel via ordinary least squares estimators (Kiebel & Holmes, 2007) and the corresponding SPM is derived with test statistic Z_v per voxel v.

To address the multiplicity arising from simultaneously testing thousands of voxels for activation, one can proceed with topological inference by testing peaks or clusters (Hayasaka & Nichols, 2004; Worsley et al., 2004). Unlike classical corrections for multiple testing such as Bonferroni, topological inference takes into account the spatial characteristics inherent to the data (Nichols, 2012).

To define peaks or clusters, two steps need to be taken. First an arbitrary threshold (denoted as u) on the SPM has to be set. From here, a search algorithm will identify the supra-threshold excursion set containing voxels with a test statistic Z_v above u. This search algorithm will either use a 6-adjacency, 18-adjacency or 26-adjacency rule to define a search region in which it will look for adjacent voxels in the excursion set. Each cluster c (with $c = 1, \ldots C$ and C the number of supra-threshold clusters in a volume) can be characterised by either the maximum test statistic (peak Z_v) within this cluster or its extent S (Figure 1). One then relies either on the parametric properties of Random Field Theory (RFT) (Worsley et al., 2002) or on the empirical properties of the permutation (PERM) null distribution (Nichols & Holmes, 2002) to form a second threshold Z_α for the identification of significant peaks or clusters while controlling the type I error rate.

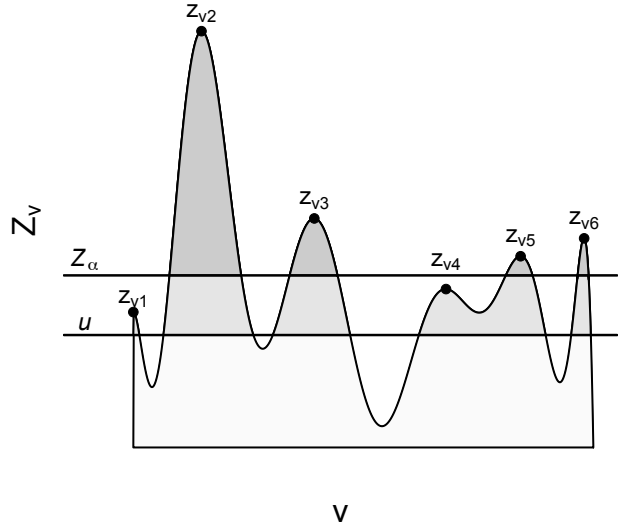


Figure 3.1 Peaks and clusters are identified after choosing a cluster-forming threshold u. A second threshold Z_α is then defined to control for the type I error rate at level α. While Z_{vc} illustrates the cluster maxima for cluster c, the darker grey areas illustrate the cluster extent S.

Both for peaks and clusters, p-values can be obtained. A peak p-value is the probability of finding under the null hypothesis of no activation a peak which is at least as high as the observed one. Analogously, a cluster p-value equals the probability of observing an equal or larger number of connected voxels when no activation is present. These p-values can be obtained in a parametric way using RFT or using a non-parametric PERM method. Using the notation from Durnez, Moerkerke, & Nichols (2014), we first derive expressions for these p-values under RFT.

Let $\Omega \subset \mathbb{R}^D$ be the D-dimensional search region of interest (i.e. the test image with $D = 3$), and $Z_v \in \mathbb{R}$ denote a random variable, i.e. the peak test statistic at voxel $v \in \mathbb{R}^D$. For parametric RFT an uncorrected p-value for a peak Z_v is approximated by

$$P(Z_v \geq z_v | Z_v \geq u, H_0) \approx exp(-u(z_v - u)) \tag{3.2}$$

with H_0, the null hypothesis of no activation. For cluster p-values, consider the spatial extent S. It is approximated by $S \approx cH^{D/2}$ (Worsley, 2007) where c equals:

$$\frac{FWHM^D u^{D/2} P(Z_v \leq u|H_0)}{EC_D(u)\Gamma(D/2+1)} \tag{3.3}$$

with $FWHM$ the Full Width Half Maximum, $EC_d(z)$ the d-dimensional Euler Characteristic density of test statistic z and Γ the gamma function. The p-value for the extent S is approximated by

$$P(S \geq s|H_0) \approx exp\left(-u\,(s/c)^{2/D}\right) \tag{3.4}$$

Although exact formulations exist (Hayasaka & Nichols, 2003), for computational convenience we opted to use the above approximation in the simulation study.

For non-parametric permutation-based inference, an empirical permutation null distribution conditional on the observed data is constructed by relying on the assumption of exchangeability of data points under the null hypothesis. If the null hypothesis of no task effect is true, the labels of an experiment are exchangeable. Hence by randomly shuffling or permuting these labels, recomputing the test statistic on different permuted data sets, an empirical null distribution can be obtained. However, following Durnez, Moerkerke, & Nichols (2014), label swapping is only applied amongst adjacent blocks to account for temporal correlations. An advantage is that the method does not rely on distributional assumptions of the test statistic but it is computationally more expensive.

Note that we only use p-values that are uncorrected for multiple testing. Although other authors prefer to derive validity via p-values that are corrected for the multiple testing problem (Hayasaka & Nichols, 2003), we use (approximated) uncorrected p-values. Among non-activated voxels, these p-values are expected to be uniformly distributed. Furthermore, the choice of the cluster-forming threshold u and the cluster-forming algorithm are known to heavily impact the results (see e.g. Petersson, Nichols, Poline, & Holmes, 1999). The latter two choices are kept constant in our study since these are not the main subject of this study (see section Simulation and Data Analysis Details).

3.2.2 Spatial Isotropic Gaussian Smoothing and Structural Adaptive Smoothing

Spatial smoothing is an essential pre-processing step in the analysis of fMRI data and aims to reduce random noise and to create a smooth contiguous image that complies with the RFT assumptions (Friston et al., 2007; Hayasaka & Nichols, 2003). By spatially smoothing raw data in one voxel, one incorporates information from the surrounding voxels. More technically, it involves replacing the BOLD signal from every voxel with a weighted sum of these values in neighbouring voxels. To this end, an isotropic 3D Gaussian kernel can be used with equally vastly decreasing weights as the distance between voxels grows (see e.g. Worsley et al., 2002). The amount of smoothing is usually expressed in a FWHM-value of the kernel (see also Figure 3.2). There is a straightforward relationship between FWHM and the standard deviation σ of a Gaussian kernel: $\text{FWHM} = 2\sqrt{2\ln 2}\sigma \approx 2.35\ \sigma$.

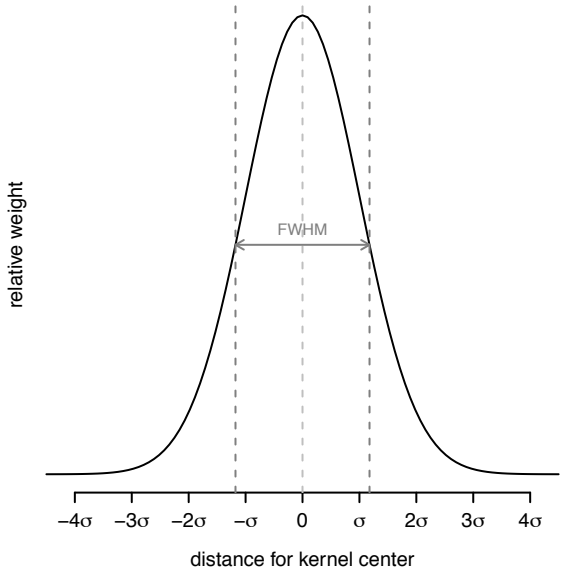


Figure 3.2 Full Width at Half Maximum (FWHM) of a Gaussian Smoothing Kernel (characterised by σ).

The potential loss of information about the spatial extent of an active area is viewed as an important shortcoming of spatial isotropic Gaussian smoothing. At the border of an activation field, activation will spread to brain areas which will falsely be declared active leading to incorrectly and poorly delineated activation regions. Tabelow et al. (2006) and Polzehl, Voss, & Tabelow (2010) address this issue by devising an adaptive weight structure. Their weight structure is set up via an iterative procedure that takes into account the activation magnitude of the test image (SPM), and thus the borders of activation. First, activity is assessed without smoothing. A minimal smoothing kernel is then used to smooth the estimates for the activation parameters. Second, the procedure checks if the estimates for the model fitting procedure improve or not. If so, the bandwidth of the smoothing kernel will increase and the procedure is repeated until the best fit is achieved or smoothing reaches a maximum specified value. When the smoothing kernel moves over non-activated fields, it will behave as its non-adaptive counterpart, but when encountering activation, it results in a better delineation of these areas.

3.2.3 Simulation and Analysis Details

Simulations are carried out in R (R Core Team, 2013) using the **neuRosim** package (Welvaert, Durnez, Moerkerke, Verdoolaeghe, & Rosseel, 2011). Brain volumes of $30 \times 30 \times 30$ voxels of 1 mm are simulated under 2 scenarios: one with no activation (null data) and one with activation in 8 small cubicles (8 times $5 \times 5 \times 5$ voxels). We opt for a blocked design of 10 blocks per condition (activated/rest), each lasting 20 seconds (upper panel in Figure 3.3). For the simulation of BOLD signals in activated voxels the canonical, double gamma, HRF (Henson & Friston, 2007) is used; for the null data, a baseline signal is used (middle panel of Figure 3.3). We set the time to repetition at 2 s. Three types of noise are added to the simulated BOLD signal: 1) white Gaussian noise ; 2) temporal $AR(1)$ noise; and 3) spatial noise based on a Gaussian random field with a FWHM of 2 (see lower panel of Figure 3.3). The contrast-to-noise ratio (CNR) is defined as the peak signal change measured from baseline (A) divided by the standard deviation of the noise in the time series: $CNR = A/\sigma_N$ with σ_N the standard deviation of the noise (Welvaert & Rosseel, 2013). In our simulations, we consider a CNR of 0.2, 0.5 and 1.0. We based these CNR-values on the estimated CNR of several subjects in the simple motor task described further on in the manuscript. Moreover, such CNR-values

are in line with e.g. the work of Churchill, Yourganov, et al. (2012). Each simulation setting is repeated 500 times.

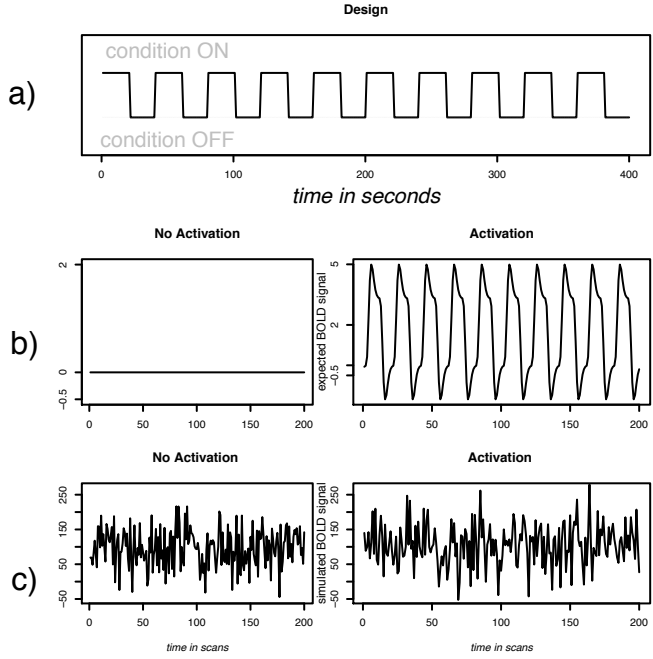


Figure 3.3 The simulation setting. a) the experimental design; b) the expected signal under no activation and activation (corresponds to the design matrix $\mathbf{X}$); c) the simulated BOLD signal under a CNR of 0.5.

Non-adaptive smoothing is performed with the **AnalyzeFMRI** package (Bordier, Dojat, & de Micheaux, 2011), while the structural adaptive smoothing is conducted using the **fmri** package (Tabelow & Polzehl, 2011). The smoothing kernel width is varied from 1 to 6 voxels for both non-adaptive and adaptive smoothing. SPMs are based on a design matrix X (see Equation 3.1) that represents the simulated signal (i.e. a convolution of the correct stimulus function with double gamma HRF, Henson & Friston, 2007). This matrix is calculated using the **fmri** package (Tabelow & Polzehl, 2011).

The cluster-defining thresholds have an upper-tail probability of 0.001 under the null of no activation (Hayasaka & Nichols, 2003). Peaks and clusters are sought with a 26-point clustering algorithm using FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). For the permutation-

based procedure, 100 permutations are performed.

We note that smoothing is expressed on the scale of voxel sizes. In our study, voxels are seen as cubicles of 1x1x1 mm. A smoothing kernel with a FWHM of 3 mm encapsulates thus a width of 3 voxels. Furthermore, for computational reasons, the full extent of the smoothing kernel was cut off at 6 times σ. As such, over the voxels within 3 times σ in each direction a re-weighting of the original signal is applied, covering approximately 99% of the surrounding information (see also Figure 3.2).

3.2.4 Evaluation of the Selection Procedure

Validity

A key characteristic of the distribution of uncorrected p-values (derived from brain null data) is its assumed uniform distribution under the null hypothesis of no activation. The more the empirical distribution matches a uniform distribution in the interval $[0, 1]$, the more valid the method. In addition to the inspection of this empirical distribution, one can assess the empirical type I error rate. The empirical type I error rate, denoted $\hat{\alpha}$, is defined here as the average proportion of spatial features per volume that is falsely declared active at a pre-specified nominal level α. To compute the empirical $\hat{\alpha}$, the average number of spatial features (peak or cluster extent) declared significant at the α-level over these simulations with no activation is divided by the average number of observed features over the simulations. Let, per simulation, m denote the number of significant clusters or peaks, and m_0 the total number of observed clusters or peaks, then

$$\hat{\alpha} = \frac{\hat{E}(m)}{\hat{E}(m_0)}. \tag{3.5}$$

$\hat{E}(m)$ is the average amount of features above the second threshold (Z_α), while $\hat{E}(m_0)$ is the average amount of features defined by cluster-forming threshold u. The closer $\hat{\alpha}$ is to α, the more valid the method.

Reliability

Reliability will be assessed in three different ways: at the voxel level, at the cluster level and at the peak level. First, we will rely on a slight modification of the Jaccard Index (Maitra, 2010; Jaccard, 1901). This index $\omega_{j,GT}$ measures the overlap between two sets, i.e. the number of active voxels (with test statistic above Z_α) from test image W_j and the

truly activate voxels W_{GT} (based on the simulated Ground Truth, GT), as follows:

$$\omega_{j,GT} = \frac{W_{j,GT}}{W_j + W_{GT} - W_{j,GT}} \tag{3.6}$$

with $W_{j,GT}$ the union of active voxels in both images. Thus, $0 < \omega_{j,l} < 1$ is the ratio of the total amount of voxels which are active in the GT and declared as active in the test image, and the number of voxels active in either the test image or the GT.

Alternatively, Gorgolewski et al. (2012) set up a trade-off (TO) measure for spatial delineation of clusters, defined as the difference between the amount of under- and overestimated voxels from a selected cluster, i.e. $TO = FP - FN$ with FP defined as the number of False Positive voxels and FN defined as the number of False Negative voxels with respect to the closest *true* cluster. Note that perfect spatial delineation only occurs when both $FP = 0$ and $FN = 0$.

As a third measure for spatial reliability, we will use the average Euclidean distance d of the estimated peak of a cluster to the true center of activation of that cluster.

Stability

Stability is related to the variability in selected features. We will assess the variability in the number of selected peaks and clusters and in cluster size. A higher variability indicates a lower stability. Moreover, the variability in the above described reliability measures will be considered as well. While the average of these reliability measures is related to the reliability, the variability of these measures is a good indicator of the stability of a method.

For an overview of all the validity, reliability and stability measures that we will consider in the simulation study, we refer to Table 3.1.

3.3 Simulation Results

Under the simulation setting described in the previous section, we explore in a factorial $2 \times 2 \times 3$ design (2 inferential methods: RFT or PERM, 2 smoothing methods: adaptive or not, and 3 smoothing kernel widths: 1, 3 or 6 voxels) the performance of the selection procedures in terms of the different reproducibility measures presented in Table 3.1. Note that throughout all tables and figures adaptive smoothing is abbreviated with “a".

Table 3.1 Measurements with their criterion and objectives to evaluate the performance of the selection procedures.

Measure	Criterion
Validity	
Distribution p-values	Uniform distribution over $[0, 1]$ under null of no activation
Type I error rate	Controlled at nominal level α
Reliability	
Overlap between test image and ground truth	High mean = high reliability
Spatial delineation of clusters (FP, FN)	Low mean = high reliability
Euclidean distance peak/centre to true centre	Low mean = high reliability
Stability	
Number of selected peaks or clusters	High variance = low stability
Cluster Size	High variance = low stability
Overlap between test image and ground truth	High variance = low stability
Spatial delineation of clusters	High variance = low stability
Euclidean distance peak/centre to true centre	High variance = low stability

3.3.1 Validity: Distribution Uncorrected p-Values and Empirical Type I Error Rate

Figure 3.4 shows the empirically obtained distribution of the p-values for inference based on cluster size (two left columns panel) or on peak height (two right columns panel) for the 12 different selection procedures. We find that p-values for cluster size do not follow the expected uniform distribution under the null, especially when the amount of smoothing is small. This holds for RFT and to a lesser extent PERM, and for both smoothing procedures. When peak based inference is used, the empirical distribution of the p-values more closely approaches the uniform distribution.

Table 3.2 shows the empirical type I error $\hat{\alpha}$ for both cluster size and peak based inference for varying values of the nominal α ($0.1, 0.05, 0.01$ and 0.001). For cluster size based inference relying on RFT, low amounts of smoothing result in liberal tests ($\hat{\alpha} > \alpha$) while too conservative type I error rates are obtained with a high amount of smoothing. For peak based inference relying on RFT, we observe the reverse pattern, namely more liberal results with increased smoothing. For RFT based inference, there is no remarkable difference between adaptive smoothing and classical Gaussian smoothing in terms of type I error. When using PERM, type I error rates close to α are obtained under all conditions.

In summary, in terms of validity, we find in our simulation study PERM to be more valid than RFT, especially for peak based inference. The main differences due to kernel width emerge in RFT based inference, while cluster size based inference relying on PERM does not depend heavily on kernel width. There are no remarkable differences between adaptive and non-adaptive smoothing.

Table 3.2 Average type I error rate $\hat{\alpha}$ with standard deviation (sd) for given α and $P(Z_v \geq u) = 0.001$ for inference based on peaks (P) and cluster-size (C) via Random Field Theory (rft) and permutation based inference ($perm$). a: adaptive smoothing, *: nominal level not within 2 standard errors from given α.

inference	smoothing	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
rft C	01	0.267 (0.097)*	0.092 (0.068)*	0.016 (0.028)*	0.001 (0.006)
perm C	01	0.089 (0.067)*	0.036 (0.041)*	0.008 (0.019)*	0.000 (0.004)*
rft P	01	0.097 (0.065)	0.044 (0.045)*	0.005 (0.015)*	0.000 (0.000)*
perm P	01	0.103 (0.067)	0.053 (0.048)	0.010 (0.022)	0.001 (0.006)
rft C	01a	1.000 (0.000)*	1.000 (0.000)*	1.000 (0.000)*	1.000 (0.000)*
perm C	01a	0.023 (0.025)*	0.023 (0.025)*	0.001 (0.004)*	0.000 (0.003)*
rft P	01a	0.095 (0.048)*	0.042 (0.032)*	0.004 (0.011)*	0.000 (0.001)*
perm P	01a	0.100 (0.049)	0.049 (0.035)	0.010 (0.016)	0.001 (0.005)
rft C	03	0.101 (0.120)	0.046 (0.075)	0.006 (0.026)*	0.000 (0.003)*
perm C	03	0.092 (0.112)	0.050 (0.081)	0.011 (0.035)	0.000 (0.000)*
rft P	03	0.121 (0.126)*	0.055 (0.087)	0.008 (0.036)	0.001 (0.011)
perm P	03	0.100 (0.118)	0.048 (0.081)	0.009 (0.036)	0.000 (0.000)*
rft C	03a	0.106 (0.109)	0.048 (0.073)	0.006 (0.023)*	0.000 (0.007)
perm C	03a	0.092 (0.105)	0.049 (0.075)	0.010 (0.033)	0.000 (0.000)*
rft P	03a	0.140 (0.127)*	0.071 (0.093)*	0.018 (0.045)*	0.002 (0.014)
perm P	03a	0.100 (0.109)	0.044 (0.076)	0.010 (0.033)	0.000 (0.000)*
rft C	06	0.074 (0.153)*	0.034 (0.102)*	0.002 (0.027)*	0.000 (0.000)*
perm C	06	0.095 (0.174)	0.051 (0.123)	0.007 (0.037)	0.000 (0.000)*
rft P	06	0.122 (0.192)*	0.061 (0.146)	0.011 (0.080)	0.001 (0.023)
perm P	06	0.093 (0.169)	0.049 (0.131)	0.012 (0.081)	0.000 (0.000)*
rft C	06a	0.066 (0.202)*	0.022 (0.115)*	0.000 (0.000)*	0.000 (0.000)*
perm C	06a	0.107 (0.245)	0.051 (0.186)	0.002 (0.027)*	0.000 (0.000)*
rft P	06a	0.182 (0.323)*	0.101 (0.246)*	0.045 (0.174)*	0.002 (0.030)
perm P	06a	0.094 (0.246)	0.050 (0.185)	0.007 (0.070)	0.000 (0.000)*

3.3.2 Reliability

Figure 3.5 shows boxplots for the adapted Jaccard Index $\omega_{j,GT}$ for a CNR of 0.2, 0.5 and 1.0. The median can be used to assess reliability of the different selection procedures. We find in general that adaptive smoothing results is superior to Gaussian smoothing, with the difference between both most pronounced with high signal. There is one notable exception, when PERM is used in combination with adaptive smoothing. Overall, there is little difference between PERM and RFT, except for the just

Table 3.3 Average number of False Positives (FP), False negatives (FN) and the Trade-off (TO) of both with standard deviation (sd) for a contrast to noise ratio (CNR) of 0.2, 0.5 and 1.0 for Random Field Theory (rft) based inference and Permutation based inference ($perm$). a: adaptive smoothing.

		CNR=0.2						CNR=0.5						CNR=1.0					
	FWHM	FN	sd	FP	sd	TO	sd	FN	sd	FP	sd	TO	sd	FN	sd	FP	sd	TO	sd
rft	01	18.84	15.37	4.6	3.45	-14.24	15.9	0.06	2.79	52.28	11.07	52.22	11.63	0	0	145.86	4.62	145.86	4.62
perm	01	17.33	8.79	4.58	3.47	-12.75	9.89	0	0.02	52.3	11.04	52.3	11.05	0	0	145.86	4.62	145.86	4.62
rft	01a	122.61	5.16	0.6	0.52	-122.02	4.98	17.77	34.55	0.71	0.86	-17.06	34.14	0.65	9.01	0.41	0.67	-0.25	8.85
perm	01a	116.82	7.27	0.19	0.51	-116.63	7.25	6.96	6.21	0.58	0.8	-6.38	6.16	0.31	6.24	0.4	0.66	0.09	6.14
rft	03	1.35	3.06	90.09	22.5	88.73	23.49	0	0	227.29	25.33	227.29	25.33	0	0	360.64	23.75	360.64	23.75
perm	03	1.29	1.32	90.11	22.48	88.82	23.15	0	0	227.29	25.33	227.29	25.33	0	0	360.64	23.75	360.64	23.75
rft	03a	20.43	19.4	16.78	15.36	-3.65	27.09	10.58	31.83	2.07	5.24	-8.51	28.25	6.75	28.24	1.43	4.45	-5.32	24.91
perm	03a	15.68	6.6	19.64	15.83	3.97	20.2	1.58	1.33	1.16	3.61	-0.42	3.76	0	0.06	0.92	3.08	0.91	3.07
rft	06	0.1	2	216.63	53.79	216.53	53.97	0	0	470.58	86.44	470.58	86.44	0	0	667.89	89.37	667.89	89.37
perm	06	0.07	0.28	216.99	53.59	216.92	53.65	0	0	470.67	86.43	470.67	86.43	0	0	668.24	89.45	668.24	89.45
rft	06a	32.96	20.18	35.13	58.59	2.17	67.51	9.16	29.29	3.05	13.2	-6.11	25.03	1.52	13.67	0.34	3.37	-1.18	11.38
perm	06a	13.65	7.74	212.7	93.07	199.05	93.05	1.12	1.13	131.62	58.39	130.5	57.9	0	0	1.96	5.79	1.96	5.79

Table 3.4 Average Euclidean distance from peak to center of activation ($\bar{d}$) with standard deviation for Random Field Theory (rft) and Permutation based inference ($perm$). a: adaptive smoothing.

	CNR=0.2				**CNR=0.5**				**CNR=1.0**			
	rft		*perm*		*rft*		*perm*		*rft*		*perm*	
FWHM	$\bar{d}$	sd	$\bar{d}$	sd	$\bar{d}$	sd	$\bar{d}$	sd	$\bar{d}$	sd	$\bar{d}$	sd
01	5.02	1.5	5.01	1.5	5.48	1.88	5.48	1.88	5.62	1.88	5.62	1.88
01a	4.8	0.24	4.88	1.02	4.92	0.85	4.91	0.86	5.05	1.19	5.05	1.19
03	4.54	1.54	4.55	1.56	4.43	1.51	4.43	1.51	4.15	1.39	4.15	1.39
03a	4.7	1.53	4.63	1.96	4.97	1.71	4.86	2.01	4.52	1.6	4.07	2.26
06	3.12	0.91	3.11	0.92	2.89	0.91	2.9	0.92	2.39	0.73	2.39	0.73
06a	2.77	0.69	1.04	1.69	2.6	0.7	0.04	0.25	2.09	0.46	0.08	2.61

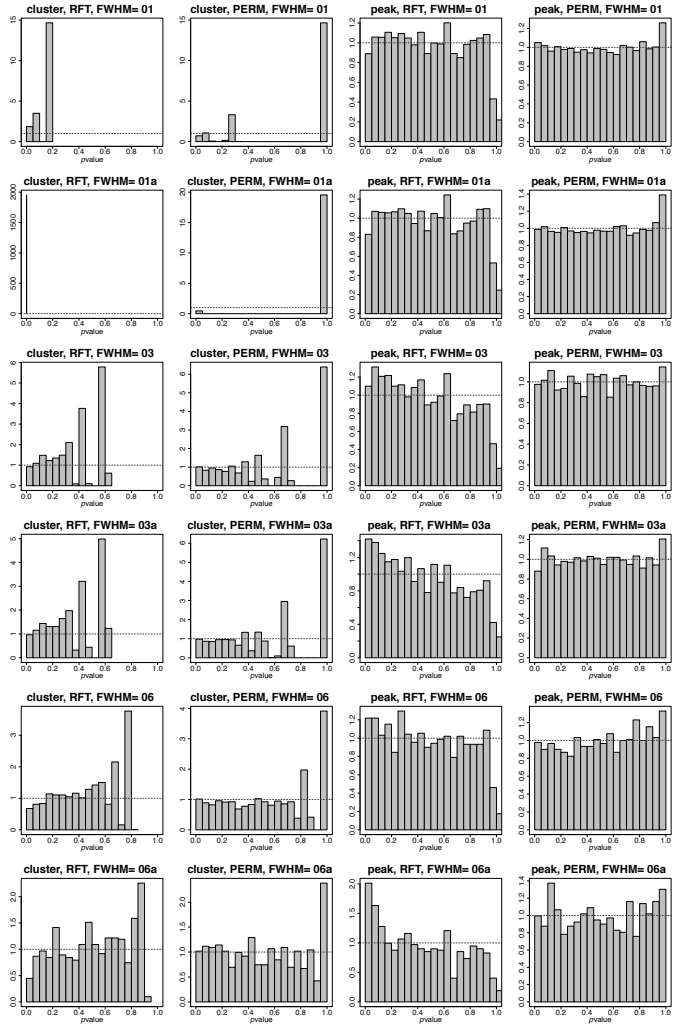


Figure 3.4 The null distribution of p-values based on random field theory (RFT) and permutation-based inference (PERM) on clusters and peaks with Z_v: a T-distributed variable with 198 degrees of freedom. $P(Z_v \geq u) = 0.001$ for FWHM of 1, 3 and 6 voxels width. a: adaptive smoothing.

mentioned combination of PERM and adaptive smoothing.

Table 3.3 shows the average false negatives (FN), false positives (FP)

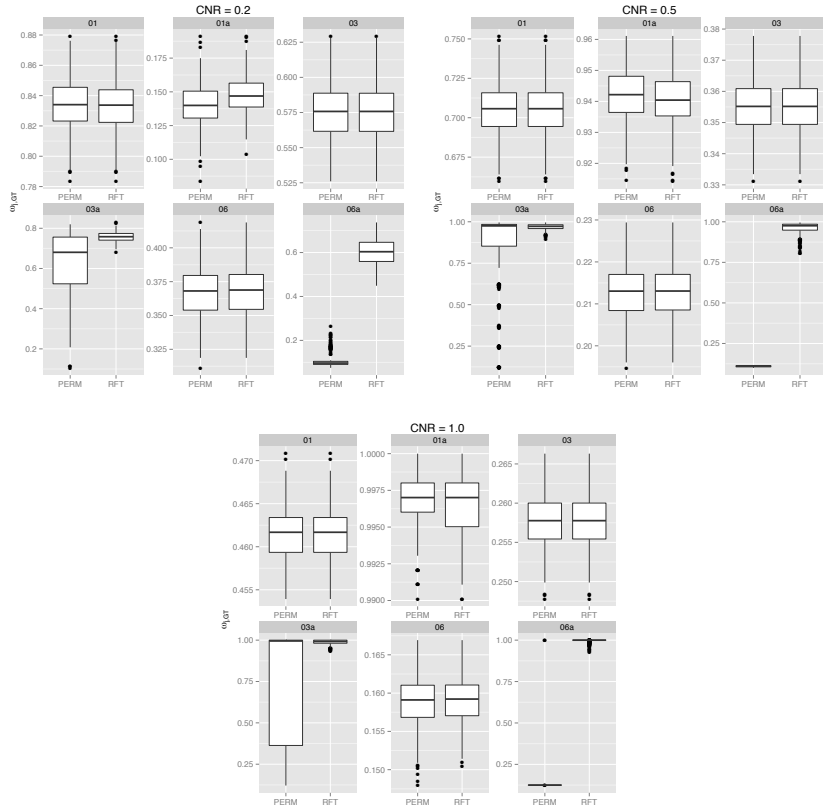


Figure 3.5 Overlap between the ground truth (GT) and test results for simulated images for Random Field Theory (RFT) and Permuation-based inference (PERM) for a contrast to noise ratio (CNR) of 0.2 (left panel), a CNR of 0.5 (middle panel) and a CNR of 1.0 (right panel). The thick line represents the median, the boxes contain 50% of the observed reliability scores. a: adaptive smoothing.

and the trade off (TO). With increasing smoothing width we typically find a reduced amount of FNs when the CNR is low. However, an increasing amount of FPs is associated with increasing kernel width. The optimal choice for selection procedure (inference, amount of smoothing and smoothing type) in terms of TO depends on the CNR though. Per cluster that is detected we find, with increasing CNR, a better delineation when adaptive smoothing is used. As already noted at the discussion of the Jaccard Index, the combination of permutation-based inference with

adaptive smoothing should be used with care.

For peak based inference, the average Euclidean distances from peak to center of activation are depicted in Table 3.4. When the amount of smoothing is high, we find on average lower distances to the center of activation. With large smoothing kernel, there is indication that adaptive smoothing results in smaller distances than classical Gaussian smoothing. No relevant differences are observed between PERM and RFT, except again in the case of adaptive smoothing with high kernel width under a high CNR, where PERM seems more reliable than RFT.

In summary, in terms of reliability, we conclude that adaptive smoothing is performing better. The amount of smoothing should be carefully considered though.

3.3.3 Stability

Table 3.5 shows the average number of selected clusters and the standard deviation on the number of selected clusters. To assess stability, we focus on the standard devation. For cluster size-based inference, we find that for lower kernel widths, PERM renders less variable results than RFT. However, with increasing kernel width, this pattern is reversed in most cases. There is also strong indication that adaptive smoothing leads to more variable results than Gaussian smoothing. In line with the reliability results, we note that adaptive smoothing and large kernel width combined with PERM results in a smaller amount of detected clusters.

The mean and standard deviation of the number of significant peaks can be found in Table 3.6. PERM results in less variable estimated number of peaks than RFT for almost all combinations, regardless the CNR. Increasing the smoothing width typically leads to less variability in the number of identified peaks.

In Table 3.7 the mean and standard deviation on the size of the selected clusters is presented. Note that true size (125 for all clusters) is best recovered when the adaptive smoothing is used, except under a low CNR and small kernel width. While under classical Gaussian smoothing, the standard deviation on the size of the selected clusters increases with increasing kernel width, for both PERM and RFT, such trend is not consistently seen for adaptive smoothing.

We can also assess stability in terms of variability of the reliability measures. With respect to variability in the spatial delineation *TO* (Table 3.3), we find again that while under classical Gaussian smoothing, this

Table 3.5 Average number ($\bar{n}$) of clusters with standard deviation (sd) for a contrast to noise ratio (CNR) of 0.2,0.5 and 1.0 for Random Field Theory (rft) based inference and Permutation based inference ($perm$). a: adaptive smoothing.

	CNR=0.2				**CNR=0.5**				**CNR=1.0**			
	rft		*perm*		*rft*		*perm*		*rft*		*perm*	
FWHM	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd
01	8.14	0.40	8.03	0.19	8.00	0.06	8.00	0.00	8.00	0.00	8.00	0.00
01a	64.08	7.59	17.21	3.39	8.83	0.94	8.02	0.13	8.04	0.21	8.02	0.14
03	8.00	0.09	8.00	0.06	7.99	0.08	7.99	0.08	7.99	0.51	7.99	0.51
03a	8.23	0.53	6.46	1.79	8.62	0.79	6.97	2.22	8.46	0.67	6.19	3.52
06	7.97	0.19	7.93	0.26	7.90	0.32	7.89	0.33	7.93	0.28	7.89	0.35
06a	8.25	0.59	1.13	0.60	8.51	0.67	1.00	0.13	8.10	0.30	2.17	0.52

Table 3.6 Average number of peaks $\bar{n}$ with standard deviation (sd) for a contrast to noise ratio (CNR) of 0.2,0.5 and 1.0 for Random Field Theory (rft) based inference and Permutation based inference ($perm$). a: adaptive smoothing.

	CNR=0.2				**CNR=0.5**				**CNR=1.0**			
	rft		*perm*		*rft*		*perm*		*rft*		*perm*	
FWHM	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd	$\bar{n}$	sd
01	19.52	2.99	18.68	2.89	15.57	2.5	14.69	2.33	15.2	2.41	14.34	2.24
01a	21.26	3.83	21.66	4.06	46.53	2.52	45.27	2.22	41.34	3.29	39.7	3.08
03	8.37	0.65	8.01	0.5	8.3	0.57	8	0	8.19	0.68	8	0.51
03a	8.6	0.78	8.05	0.7	8.56	0.77	8.01	0.09	8.4	0.65	8.01	0.12
06	8.14	0.38	8	0.36	8.1	0.32	8	0	8.04	0.19	8	0
06a	8.17	0.4	8.01	0.12	8.09	0.3	8.02	0.13	8.05	0.21	8.03	0.18

Table 3.7 Average size of clusters $\bar{S}$ with standard deviation (sd) for a contrast to noise ratio (CNR) of 0.2,0.5 and 1.0 for Random Field Theory (rft) based inference and Permutation based inference ($perm$). a: adaptive smoothing.

	CNR=0.2				**CNR=0.5**				**CNR=1.0**			
	rft		*perm*		*rft*		*perm*		*rft*		*perm*	
FWHM	$\bar{s}$	sd	$\bar{s}$	sd	$\bar{s}$	sd	$\bar{s}$	sd	$\bar{s}$	sd	$\bar{s}$	sd
01	110.76	15.9	112.25	9.89	177.22	11.63	177.3	11.05	270.86	4.62	270.86	4.62
01a	2.98	4.98	8.37	7.25	107.94	34.14	118.62	6.16	124.75	8.85	125.09	6.14
03	213.73	23.49	213.82	23.15	352.29	25.33	352.29	25.33	485.64	23.75	485.64	23.75
03a	121.35	27.09	128.97	20.2	116.49	28.25	124.58	3.76	119.68	24.91	125.91	3.07
06	341.53	53.97	341.92	53.65	595.58	86.44	595.67	86.43	792.89	89.37	793.24	89.45
06a	127.17	67.51	324.05	93.05	118.89	25.03	255.5	57.9	123.82	11.38	126.96	5.79

variability increases with increasing kernel width, such trend is less clear for adaptive smoothing. Finally, we find that when adaptive smoothing is applied, PERM typically yields more variable estimated distances to the cluster center than RFT (Table 3.4).

3.4 Assessment of Stability in Real Data

We restrict the illustration to the demonstration of the assessment of the data analytical stability on real data. Unlike simulated data, the ground truth is unknown in real data, which complicates the assessment of validity and reliability. Examples on the assessment of validity and reliability can be found elsewhere (Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013; Maitra, 2010; Rombouts, Barkhof, Hoogenraad, Sprenger, & Scheltens, 1998). The illustration here uses data of one subject in a simple motor task (Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013; Gorgolewski, Storkey, Bastin, Whittle, Wardlaw, & Pernet, 2013). In the experiment subjects had to conduct three simple motor tasks: movement of the feet, movement of the lips or movement of the fingers. Here we focus on the specific contrast that compares the expected activation during movement of the foot with expected the movement of the lips and fingers. The estimated strength of the activation (expressed by the CNR) in the activated areas equals 1.23.

While in a simulation setting, one can repeatedly draw samples from the true underlying distribution, it is more complex to obtain repeated samples from real data. One can for example rely on bootstrapping techniques in which GLM residuals are resampled, after temporal decorrelation based on a parametric form of the temporal variance-covariance structure (Friman & Westin, 2005). Friman & Westin (2005) explicitly relied on a parameterisation of the temporal noise structure to explore the performance of the bootstrap in the GLM-framework. However, such assumption is not necessarily needed per se. Taking into account the dependency by blocking groups of consecutive observations (hereafter referred to as the semi-parametric approach) may also offer a useful alternative (Davison & Hinkley, 1997; Lahiri, 2003). It only requires the specification of the mean-structure but excludes the need for the explicit parametrisation of a complex noise structure.

In order to resample the data here, we used such semi-parametric blocked bootstrap scheme and opted for a moderate block length of 9 consecutive time points in a moving block bootstrap scheme (Lahiri, 2003).

In the moving block bootstrap, data are split into $n - b + 1$ overlapping blocks of length b. Observation 1 to b will be block 1, observation 2 to $b+1$ will be block 2 etc. Then from these $n - b + 1$ blocks, n/b blocks will be drawn at random with replacement. Given the computational intensity of the whole procedure, we resampled only 100 times. In the pre-processing phase, we applied a temporal high-pass filter with a cut-off of 100 s, while the smoothing kernel was varied between 4, 8 and 12 mm width (i.e. the most frequently used settings in fMRI research according to Carp, 2012) ($\sim$ 1, 2 and 3 voxels width kernel). As in the simulation setting, we both applied adaptive and non-adaptive smoothing. For the inferential phase, we both assessed RFT and PERM. Stability is assessed using the variability in the number of selected clusters and peaks as a criterion.

Results are presented in Table 3.8. For the stability of the number of clusters, we find that under classical smoothing, RFT performs better than PERM, while the trend is reversed for adaptative smoothing (except for the largest smoothing width). For the stability of the number of peaks, the smoothing width and kernel type play a role. Under low smoothing, the stability for PERM is better than for RFT, and adaptative better than classical smoothing. On the contrary, under high smoothing, the stability for PERM is worse than for RFT for classical Gaussian smoothing, while for adaptive smoothing we find PERM to be less variable for adaptive smoothing.

We also find interesting correspondences between the simulation study and the real data example. Consider cluster-based inference with 3 voxels width smoothing ($\sim$12 mm in real example) and peak-based inference with 1-3 voxels kernel width ($\sim$ 8-12 mm). Based on the simulations these are the cases that result in high validity. We find that for cluster extent based inference relying on permutation methods, broad adaptive smoothing evokes additional variability over replication samples. When peak-based inference is used however under the same conditions, less variability is observed. As mentioned, peakwise inference using random field theory seemed also less problematic in terms of validity and may therefore be preferred over inference for cluster extent.

3.5 Discussion

Today's neuroimaging research goes beyond the pure development of methods and increased attention is paid to the consequences of such methods in terms of reliability. Recent contributions about power in fMRI stud-

Table 3.8 Average number of clusters and peaks ($\bar{n}_c$ and $\bar{n}_p$) with standard deviation (sd) for inference based on Random Field Theory (rft) and Permutation methods ($perm$) for the real dataset from Gorgolewski, Storkey, Bastin, Whittle, Wardlaw, & Pernet (2013). a: adaptive smoothing.

	rft		**perm**		**rft**		**perm**	
smoothing	$\bar{n}_c$	sd	$\bar{n}_c$	sd	$\bar{n}_p$	sd	$\bar{n}_p$	sd
04	1.54	0.66	7.09	3.10	21.80	9.45	19.08	8.31
04a	2.19	0.54	1.33	0.12	7.96	3.54	4.33	1.92
08	1.00	0.09	1.79	0.71	2.92	1.08	3.85	1.88
08a	1.33	0.17	2.00	0.10	7.94	3.61	4.65	1.95
12	1.00	0.10	2.05	0.89	3.25	1.24	3.97	1.77
12a	1.00	0.11	3.03	1.33	20.67	8.74	12.16	5.26

ies (Button et al., 2013; Wilke, 2012), about reliability of fMRI methods (Bennett & Miller, 2010, 2013) and about the development of techniques to these ends (see e.g. Shou et al., 2013; Maitra, 2010) stress the importance of reproducibility. Until now, these lines of research mainly focused on reliability as a proxy for reproducibility in a test-retest settings (Gorgolewski et al., 2012; Rombouts et al., 1998). Our investigation adds to the urging necessity of a stringent verification of the impact of methodological choices. Indeed, we extended the classical evaluation protocol for methodological choices leading to the selection of features. Clearly distinguishing between validity, reliability and stability, we set up a list of criteria to evaluate selection procedures.

The goal of our proposed framework is similar to that of the NPAIRS framework, introduced by Strother et al. (2002): enhancing reproducible results. While the NPAIRS framework also aims at optimizing and evaluating various elements in the selection mechanism (pipelines) in fMRI data analysis (Strother et al., 2004), we present an additional perspective on reproducibility by introducing the concept of data analytical stability in the evaluation of methods. This differentiation is lacking in the NPAIRS framework, where reproducibility is quantified via the overlap between images in a resampling context (i.e. comparable to what we refer to as 'reliability'). The differentiation between concepts of reproducibility allows the end-user to balance any evaluation according to his/her own needs. For example, in a pre-surgical setup, false-positives (falsely indicating regions as active, and as such leading to prevention - if possible - to remove the tissue) might be considered less problematic while regions

declared as active should vary as little as possible over replications. In experimental neuroimaging studies in psychology, it may be more preferable to have methods that result in peak activation that show up consistently ($\sim$ stability) rather than to have exact peaks localisation ($\sim$ reliability). Such clear distinction between concepts should aid the evaluation.

We illustrated such evaluation for two competing approaches to smoothing and to inference based either on local maxima or on cluster extent. We acknowledge that the presented evaluation does not cover the entire selection procedure, but it lays out a data analytical way to assess stability that practitioners can use making their preferred choices. The current study is as such a balance between exhaustiveness (not selecting all elements from the selection procedure) and specificity (the particular choice for cluster-based inference and its associated assumptions on smoothness). Our conjecture is though that phase 2 and 4 may matter most (a statement that needs further investigation). Evidently, we could have made other choices in each of these phases too, for instance using Independent Component Analysis or PCA driven denoising as a pre-processing step (see e.g. Churchill, Oder, et al., 2012; Churchill, Yourganov, et al., 2012). Also rather than GLM using an empirical HRF other alternatives for modeling could have been evaluated (see e.g. Afshin-Pour, Hossein-Zadeh, Strother, & Soltanian-Zadeh, 2012; Zhang et al., 2009). The concepts that were laid out in this paper however are not confined to a particular method and in principle broadly applicable. As a minimal result, we at least found evidence that choices in those phases matter substantially.

For the specific methods we evaluated and under the assumed data generating mechanisms, we showed that while typically the validity of inference methods is inferred from the empirical type I error, the inspection of the empirical distribution of p-values allows us to detect deviations that jeopardise the uniformity and as such the validity of methods. For permutation-based inference, p-values were found to vary over the entire interval $[0; 1]$. By contrast, this was not the case for random field theory based inference for cluster extent, indicating more severe distortion of the latter, which is in line with recent findings (Durnez, Roels, & Moerkerke, 2014). For random field theory based inference, we found the degree of smoothing to impact the validity: while permutation-based inference did not vary with varying smoothing kernels, random field theory based inference is characterised with fluctuating type I error rates, both for peak and cluster extent based inference. We found smoothing kernel width to be a key determinant for reliability. In line with the initial work of Tabelow et

al. (2006), we found a better trade off between false negatives and false positive for the adaptive smoothing with contrasted with non-adaptive smoothing.

By nature, data analytical stability allows to distinguish between competing approaches in the selection procedure for significant features. Given the current plentitude of selection procedures to analyse fMRI data (Carp, 2012) and the lack of an exclusive golden standard to analyze them, data analytical stability is a clear asset for such relative comparison. In addition to its use as a relative measure, data analytical stability also readily allows to provide an absolute indication of reproducibility within any study. This can be achieved for example by setting up a 95 % stability interval for the number of detected features. The latter could be formed by the 2.5th and 97.5th percentile of the number of detected features in the resampled samples, but may highly depend on the smoothness of the volume for example. Incorporation of data analytical stability at the voxel level might be more informative but requires further research (Durnez, Roels, & Moerkerke, 2014).

Stability was evaluated both in a simulation setting as well as on real data. By considering a wide range of CNR-values in our simulation study mimicking those observed in the real data example, we hope to have covered a wide range of plausible scenarios in realistic settings, but we acknowledge the arbitrariness of such choices in simulation settings. Our aim was not to provide advice on specific methods but rather on how to assess performance in particular settings with respect to properties that are relevant. The application on real data confirmed the need to acknowledge the variability resulting from choices in the selection procedure. Also, while in this study we focused on an application on the single-subject context, the concept can easily be generalized to second-level analyses too. It is interesting to note that while test-retest measures are complicated by the (often untestable) assumption of identical measurements on both test and re-test, assessment of data analytical stability as proposed here does not rely on such an assumption. Our hope is that the concept of data analytical stability will prove its use in the neuroimaging field as much as it has already done in statistical genetics (see e.g. Gordon, Glazko, Qiu, & Yakovlev, 2007; Gordon, Chen, Glazko, & Yakovlev, 2009).

Conclusion

Bennett & Miller (2013) recently raised the difficult issue of quantifying reproducibility. These authors referred to reproducibility as a *quali-*

tative measure of the ability to obtain similar results over time (Bennett & Miller, 2013, p. 1). In this work we demonstrated however that quantitative approaches to reproducibility by focusing on stability are possible and can be complementary to reliability and validity.

Acknowledgement

The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI.

References

Afshin-Pour, B., Hossein-Zadeh, G.-A., Strother, S. C., & Soltanian-Zadeh, H. (2012). Enhancing reproducibility of fMRI statistical maps using generalized canonical correlation analysis in NPAIRS framework. *NeuroImage*, *60*(4), 1970–81.

Ashby, F. G. (2011). *Statistical analysis of fmri data*. Cambridge, MA: MIT Press.

Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, *62*(2), 891–901.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–55.

Bennett, C. M., & Miller, M. B. (2013). fmri reliability: Influences of task and experimental design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690–702.

Bordier, C., Dojat, M., & de Micheaux, P. L. (2011). Temporal and spatial independent component analysis for fmri data sets embedded in the AnalyzeFMRI R package. *Journal of Statistical Software*, *44*(9), 1–24.

Button, K. S., Ioannidis, J. P. a., Mokrysz, C., Nosek, B. a., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, *14*(5), 365–76.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., ... Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Human brain mapping*, *33*(3), 609–27.

Churchill, N. W., Yourganov, G., Oder, A., Tam, F., Graham, S. J., & Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. *PloS one*, *7*(2), e31147.

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge: University Press.

Della-Maggiore, V., Chau, W., Peres-Neto, P. R., & McIntosh, A. R. (2002). An Empirical Comparison of SPM Preprocessing Parameters to the Analysis of fMRI Data. *NeuroImage*, *17*(1), 19–28.

Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, *84*, 45–64.

Durnez, J., Roels, S., & Moerkerke, B. (2014). Multiple testing in fmri: a case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, *56*(4).

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Friman, O., & Westin, F.-J. (2005). Resampling fmri time series. *NeuroImage*, *25*, 859-867.

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T., & Penny, W. (Eds.). (2007). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier Ltd./Academic Press.

Friston, K. J., Holmes, A., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical Parametric Maps in Functional Imaging: A general Linear Approach. *Human Brain Mapping*, *2*, 189–210.

Gordon, A., Chen, L., Glazko, G., & Yakovlev, A. (2009). Balancing Type One and Two Errors in Multiple Testing for Differential Expression of Genes. *Computational statistics & data analysis*, *53*(5), 1622–1629.

Gordon, A., Glazko, G., Qiu, X., & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, *1*(1), 179–190.

Gorgolewski, K. J., Storkey, A., Bastin, M. E., Whittle, I. R., Wardlaw, J. M., & Pernet, C. R. (2013). A test-retest fMRI dataset for motor, language and spatial attention functions. *GigaScience*, *2*(1), 6.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, *6*, 1–14.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject fMRI test-retest reliability metrics and confounding factors. *NeuroImage*, *69*, 231–43.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Hayasaka, S., & Nichols, T. E. (2004). Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, *23*(1), 54–63.

Hayasaka, S., Phan, K. L., Liberzon, I., Worsley, K. J., & Nichols, T. E. (2004). Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, *22*(2), 676–87.

Henson, R., & Friston, K. J. (2007). Convolution models for fmri. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 193-210). Elsevier Ltd./Academic Press.

Jaccard, P. (1901). Distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547-579.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Kiebel, S., & Holmes, A. P. (2007). The general linear model. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer-Verlag, Inc.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464.

Maitra, R. (2010). A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. *NeuroImage*, *50*(1), 124–35.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–5.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Petersson, K. M., Nichols, T. E., Poline, J. B., & Holmes, a. P. (1999). Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *354*(1387), 1261–81.

Poline, J.-B., & Brett, M. (2012). The general linear model and fMRI: Does love last forever? *NeuroImage*, *62*(2), 871–880.

Polzehl, J., Voss, H. U., & Tabelow, K. (2010). Structural adaptive segmentation for statistical parametric mapping. *NeuroImage*, *52*(2), 515–23.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, *7*.

R Core Team, . (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rombouts, S., Barkhof, F., Hoogenraad, F. G., Sprenger, M., & Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magnetic resonance imaging*, *16*(2), 105–13.

Shou, H., Eloyan, a., Lee, S., Zipunnikov, V., Crainiceanu, a. N., Nebel, M. B., ... Crainiceanu, C. M. (2013). Quantifying the reliability of image replication studies: The image intraclass correlation coefficient (I2C2). *Cognitive, affective & behavioral neuroscience*.

Smith, S. M., Jenkinson, M., Beckmann, C., Miller, K., & Woolrich, M. (2007). Meaningful design and contrast estimability in fmri. *NeuroImage*, *34*, 127-136.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., ... Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, *15*(4), 747–71.

Strother, S. C., La Conte, S., Kai Hansen, L., Anderson, J., Zhang, J., Pulapura, S., & Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage*, *23 Suppl 1*, S196–207.

Tabelow, K., & Polzehl, J. (2011). Statistical Parametric Maps for Functional MRI Experiments in R : The Package fmri. *Journal of Statistical Software*, *44*(11).

Tabelow, K., Polzehl, J., Voss, H. U., & Spokoiny, V. (2006). Analyzing fMRI experiments with structural adaptive smoothing procedures. *NeuroImage*, *33*(1), 55–62.

Viviani, R., Grön, G., & Spitzer, M. (2005). Functional principal component analysis of fMRI data. *Human brain mapping*, *24*(2), 109–29.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaeghe, G., & Rosseel, Y. (2011). Journal of Statistical Software. *Journal of Statistical Software*, *44*(10), 10: 1-18.

Welvaert, M., & Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for FMRI data. *PloS one*, *8*(11), e77089.

Wilke, M. (2012). An iterative jackknife approach for assessing reliability and power of FMRI group analyses. *PloS one*, *7*(4), e35578.

Worsley, K. J. (2007). Random field theory. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 232-236). Elsevier Ltd./Academic Press.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*(1), 1–15.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A Unified Statistical Approach for Determining Significant Signals in Images of Cerebral Activation. *Human Brain Mapping*, *73*.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *NeuroImage*, *23 Suppl 1*, S189–95.

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., & Strother, S. C. (2009). Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magnetic resonance imaging*, *27*(2), 264–78.

4 Evaluation of Second-Level Inference in fMRI Analysis

Abstract We investigate the impact of decisions in the second-level (i.e. over subjects) inferential process in functional Magnetic Resonance Imaging (fMRI) on 1) the balance between false positives and false negatives and on 2) the data analytical stability, both proxies for the reproducibility of results. Second-level analysis based on a mass univariate approach typically consists of 3 phases. First, one proceeds via a general linear model for a test image that consists of pooled information from different subjects (Beckmann, Jenkinson, & Smith, 2003). We evaluate models that take into account first-level (within-subjects) variability and models that do not take into account this variability. Second, one proceeds via permutation-based inference or via inference based on parametrical assumptions (A. P. Holmes, Blair, Watson, & Ford, 1996). Third, we evaluate 3 commonly used procedures to address the multiple testing problem: familywise error rate correction, false discovery rate correction and a two-step procedure with minimal cluster size (Lieberman & Cunningham, 2009; Bennett, Wolford, & Miller, 2009). Based on a simulation study and on real data we find that the two-step procedure with minimal cluster-size results in most stable results, followed by the familywise error rate correction. The false discovery rate results in most variable results, both for permutation-based inference and parametrical inference. Modeling the subject-specific variability yields a better balance between false positives and false negatives when using parametric inference.

This chapter has been published in Computational Intelligence and Neuroscience.
Roels, S. P., Loeys, T., & Moerkerke, B. (2016). Evaluation of Second-Level Inference in fMRI Analysis. *Computational Intelligence and Neuroscience, 2016*, Article ID 1068434, 22 pages.

4.1 Introduction

In cognitive neurosciences, functional Magnetic Resonance Imaging (fMRI) plays an important role to localize brain regions and to study interactions among those regions (respectively functional segregation and functional integration, see e.g. Friston, 2007) The analysis of an fMRI time course in a single subject (first-level analysis) offers some insight into subject-specific brain functioning while group studies that aggregate results over individuals (second-level analysis) yield more generalizable results. In this paper, we focus on the mass univariate approach in which the brain is divided in small volume units or voxels, although alternatives exist (e.g. Vahdat, Maneshi, Grova, Gotman, & Milner, 2012). For each of these voxels, a general linear model (GLM) is used to model brain activation, at the first and the second level (Lindquist, 2008). The activation is then judged at the voxel level, rather than based on topological features. The selection of activated voxels can be viewed as a sequence of different phases (Roels, Bossier, Loeys, & Moerkerke, 2015). For first-level analyses, Carp (2012) demonstrated the large variation in the choices made in each of these different phases which impacts results. In second-level analyses, although to a lesser extent, different combinations of choices are possible too. We consider the following phases in the analysis of group studies: (1) aggregation of data over subjects, (2) inference and (3) correction for multiple testing.

In two commonly used software programs to analyze fMRI data (i.e. SPM and FSL Carp, 2012), the expected activation in each voxel is modeled in a two-step approach (Beckmann et al., 2003). In the first-level analysis, the evidence per subject is summarized in a linear contrast of the parameters, necessary to model the study design. These contrast images are then passed to the second-level analysis in which the evidence is weighted over subjects. To pool this information over subjects, one can either take into account subject-specific variability in constructing the voxel-wise test statistics or only rely on the estimated contrasts and not take into account this subject-specific variability (Mumford & Nichols, 2006).

After pooling the data, one proceeds to the second phase, the inference phase. While parametric inference offers the advantage of closed-form null distributions that can be used to obtain p-values, it depends on strong assumptions which are not easy to satisfy in practice (Nichols & Hayasaka, 2003) and have not been tested extensively (Nichols, 2012). An alternative is to use non-parametric methods such as permutation-based inference to

create an empirical null distribution conditional on the observed sample (A. P. Holmes et al., 1996; Nichols & Holmes, 2002; Nichols, 2012).

Third, inference must be corrected for the huge multiple testing that is induced by the mass univariate approach in which simultaneously over 100 000 tests are performed. As Bennett et al. (2009) and Lieberman & Cunningham (2009) discuss, there was (and yet is) no golden standard to address the choice for multiple testing corrections . We consider three different multiple testing procedures: controlling the False Discovery Rate (FDR), controlling the Familywise Error rate (FWE) and an approach based on uncorrected testing combined with a minimal cluster size. While FDR (Benjamini & Hochberg, 1995; Genovese, Lazar, & Nichols, 2002) and FWE control (see e.g. Nichols & Hayasaka, 2003) have a strong theoretical background with a focus respectively on the proportion of false positives among all selected voxels and on the probability to observe at least one false positive, the third approach is purely empirical in nature (Lieberman & Cunningham, 2009).

These three corrections are designed to control the multiple testing problem at the voxel level. Other popular alternatives that focus on topological features such as cluster size (i.e. the size of a neighboring collection of voxels) or cluster height exist as well. In a recent study, Woo, Krishnan, & Wager (2014) advocates against the use of cluster-based inference and demonstrate its problematic use when studies are sufficiently powered. By definition, it is cumbersome to interpret the findings resulting from "significant clusters" because these may not reflect a set of significant constituting voxels (see also Nichols, 2012). On the other hand, the third approach (Lieberman & Cunningham, 2009) resembles cluster-based testing but instead of setting a threshold for cluster size based on cluster significance, a fixed pre-specified threshold for the minimum cluster size is set. For completeness, we therefore also extend the third approach by choosing the threshold as in cluster-based inference. However, it is important to point out that we do not intend to investigate cluster-based testing which is fundamentally different from the approach taken here and relies on different topological assumptions. Instead, we focus on voxel-wise testing (for an elaborate investigation of cluster-based testing we refer to Roels, Bossier, et al., 2015).

The choices made in each of the 3 phases of a second-level analysis are crucial steps in the analysis of fMRI data and may consequently influence results. The use such second-level analyses or group studies is widespread (A. Holmes & Friston, 1998; Mumford & Nichols, 2009; Beckmann et al.,

2003; A. P. Holmes et al., 1996) but the impact of varying procedures at the different phases has not yet been extensively validated. One can distinguish three different aspects in the evaluation of methods (Roels, Bossier, et al., 2015): validity, reliability and stability. The validity can be assessed by verifying whether the false positive rate is controlled at a pre-defined, nominal level. Further, the balance between Type I errors (false positives) and Type II errors (false negatives) has long been the main interest in the validation of testing procedures (e.g. Nichols & Hayasaka, 2003). One has also acknowledged the importance of investigating the reliability of methods (e.g. Wilke, 2012; Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013). The extent to which a method is reliable can be measured through the overlap between activated brain regions over repeated measures, for example in test-retest settings.

The concept of data analytical stability, originally developed in genetics (Qiu, Xiao, Gordon, & Yakovlev, 2006), was recently introduced into the context of fMRI data analysis (Roels, Bossier, et al., 2015). This measure allows to quantify reproducibility of results through the variability on different measures, for example the variance on the number of selected voxels over replications (either in simulation studies with a known ground truth or through subsampling of real data). Stable methods are characterized by a low variability on the number of selected voxels. Data analytical stability is thus a useful additional criterion to distinguish between methods. In this paper, we assess the influence of different choices made in the three phases on the reproducibility of results. We hereby focus on the balance between false positives and false negatives and on the stability as measures for reproducibility.

In section 4.2 we give a brief overview of the different techniques. Next, we describe the details and the results of our simulation study. In section 4.4, the results and the details from the real data application. In the Discussion, we summarize our findings and end with some recommendations for the practitioner.

4.2 Methods

In this section we provide an overview on the different inferential techniques that we will consider in the simulation study and real data example. First, we describe the methods for pooling the evidence over subjects in the mass univariate GLM approach for fMRI data at the second level. Next, we summarize different multiple testing strategies that are fre-

quently exploited in the fMRI literature, such as approaches that control the familywise error rate, approaches for control of the false discovery rate and a two-step procedure based on an uncorrected threshold but requiring a minimum cluster size. Finally, we discuss the construction of test statistics under the null hypothesis that rely on parametric assumptions versus non-parametric approaches.

4.2.1 Voxel-based GLM Approach to Analyze fMRI Data at the Group Level

Group-level inference typically proceeds via a two-step procedure (Beckmann et al., 2003). In the first step, an analysis is conducted at the voxel level for each subject m separately (with $m = 1, \ldots, M$), and an appropriate contrast of interest is constructed. In a second step, these contrast images are combined to weight evidence over the M subjects.

First-level analysis

For each subject m, the BOLD signal is sampled on T time points in every voxel v (with $v = 1, \ldots, V$) during an fMRI experiment. For every voxel v, a general linear model (GLM) is then used to relate the voxels' time course (i.e. the BOLD signal) $\mathbf{Y_v} = (Y_{v1}, \ldots Y_{vt}, \ldots Y_{vT})$ to the expected BOLD signal under brain activation in the experimental setup (the design matrix $\mathbf{X}$) (see e.g. Kiebel & Holmes, 2007; Poline & Brett, 2012; Friston et al., 1995; Worsley et al., 2002):

$$\mathbf{Y_v} = \mathbf{X}\boldsymbol{\beta_v} + \boldsymbol{\varepsilon_v} \tag{4.1}$$

The design matrix $\mathbf{X}$ is the product of a convolution of the stimulus onset function with a hemodynamic response function (HRF) (e.g. Henson & Friston, 2007). When fitting Model (4.1), one needs to account for the residual correlation between consecutive time points. Let $\mathbf{A}\sigma_\varepsilon^2$ represent the variance-covariance matrix of $\boldsymbol{\varepsilon}_v$ in Model (4.1). To deal with the temporal correlation, a matrix $\boldsymbol{\Sigma_d}$ is typically constructed such that $\boldsymbol{\Sigma_d A \Sigma_d^t} = \mathbf{I}$ holds. If $\mathbf{A}$ and $\mathbf{X}$ are correctly specified, $\boldsymbol{\beta_v}$ can be unbiasedly estimated via a simple least squares approach. By relying on 'de-correlated' or whitened outcome and predictor, i.e. $\mathbf{Y}$ and $\mathbf{X}$ are pre-multiplied by $\boldsymbol{\Sigma_d^{-1}}$, an unbiased estimator for the variance of the estimator for $\boldsymbol{\beta_v}$ is obtained (see e.g. Lindquist, 2008; Cochrane & Orcutt, 1949; Kutner, Nachtsheim, Neter, & Li, 2005). Testing for specific differ-

ences between the activation in conditions for voxel v is then possible by testing the appropriate contrasts of the elements of $\boldsymbol{\beta}_v$ with a contrast vector $\mathbf{c}$, i.e. test $H_0: \mathbf{c}\boldsymbol{\beta}_v = 0$.

Second-level analysis

Next we focus on the group level analysis for a specific voxel v ($v = 1, \ldots, V$). For ease of notation, we will drop the voxel index v in the text below. For the contrast of interest, let $\boldsymbol{b} = [b_1, \ldots, b_M]^t$ denote $\left[\boldsymbol{c\hat{\beta}_1}, \ldots, \boldsymbol{c\hat{\beta}_M}\right]^t$, the estimated contrasts at the first level for subjects 1 to M. Obviously, those contrasts are not exactly known, but estimated with some imprecision. Suppose for now that those contrasts are known, and denoted by $\boldsymbol{c\beta}$, then a GLM can be used to weight the group evidence (e.g. Mumford & Nichols, 2009):

$$\boldsymbol{c\beta} = \boldsymbol{X_M\gamma} + \boldsymbol{\eta}, \tag{4.2}$$

where $\boldsymbol{X_M}$ denotes the design matrix. In the simplest case where one is interested in knowing whether there is activation over all subjects, the design matrix $\boldsymbol{X_M}$ equals a simple column matrix consisting of M elements 1. Alternatively, in the presence of between-subjects conditions or groups (for example one wants to know whether the activation is different between males and females), $\boldsymbol{X_M}$ can take more complex forms with additional regressors. Furthermore $\boldsymbol{\eta}$ is the group error vector, with $\mathrm{Var}(\boldsymbol{\eta}) = \sigma_\eta^2 \boldsymbol{I_M}$ with $\boldsymbol{I_M}$ the identity matrix of dimension M and σ_η^2 the between-subject variance.

In practice however $\boldsymbol{c\beta}$ is unknown, and instead $\boldsymbol{b}$ is used as outcome

$$\boldsymbol{b} = \boldsymbol{X_M\gamma} + \boldsymbol{\eta^*}, \tag{4.3}$$

with $\boldsymbol{\eta^*} = [\eta_1^*, \ldots, \eta_M^*]^t$ and $\boldsymbol{\eta^*} \sim N\left(0, \boldsymbol{\Sigma_\eta^*}\right)$. Since $\boldsymbol{\eta^*} = \boldsymbol{c\beta} - \boldsymbol{b} + \boldsymbol{\eta}$, it follows that the variance-covariance matrix Σ_η^* consists of the sum of two

parts:

$$\boldsymbol{\Sigma}^{*}_{\boldsymbol{\eta}} = \mathrm{var}_M(\boldsymbol{b}) + \sigma^2_{\eta}\boldsymbol{I_M} \tag{4.4}$$

$$\boldsymbol{\Sigma}^{*}_{\boldsymbol{\eta}} = \boldsymbol{\Sigma_M} + \sigma^2_{\eta}\boldsymbol{I_M} \tag{4.5}$$

$$\boldsymbol{\Sigma}^{*}_{\boldsymbol{\eta}} = \underbrace{\begin{bmatrix} \sigma^2_1 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma^2_M \end{bmatrix}}_{\text{within-subject}} + \underbrace{\sigma^2_{\eta}\boldsymbol{I_M}}_{\text{between-subject}} \tag{4.6}$$

The first term in the right hand side of (4.4) is inherent to the uncertainty associated with the estimation of $\boldsymbol{c\beta_m}$, the within-subject variability; while the second term is related the variability in the estimation of $\boldsymbol{\gamma}$, i.e. the between-subjects variance.

In the literature on multi-subject fMRI data analysis, two ways of dealing Σ_m are frequently used. Below, we refer to these two approaches as the ordinary least squares (OLS) approach and the weighted least squares (WLS) approach, respectively.

OLS: the homoscedastic case In the first case, described in A. Holmes & Friston (1998), one assumes that within-subject variances do not differ over subjects and that the residual noise is homogeneous across all M subjects. Assuming that $\sigma^2_1 = \ldots = \sigma^2_M$ simplifies the form of Σ^*_η (in Model (4.6)) to

$$\boldsymbol{\Sigma_{\eta *}} = \sigma^2_{\text{OLS}}\boldsymbol{I_M}. \tag{4.7}$$

This implies that the within- and between-subject variability cannot be disentangled.

Mumford & Nichols (2009) demonstrate that $\boldsymbol{\gamma}$ in model (4.3) (p 1470, in Equation (6)) can then be estimated as $\hat{\boldsymbol{\gamma}}_{\text{OLS}} = \boldsymbol{X_m^{-1}b}$ while the residual error variance σ^2_{OLS} is estimated as $\left(\boldsymbol{b} - \boldsymbol{X_m}\hat{\boldsymbol{\gamma}}\right)'\left(\boldsymbol{b} - \boldsymbol{X_m}\hat{\boldsymbol{\gamma}}\right)/(M-1)$). Hence, this simply amounts to solving the normal equations in the simple linear regression case and inference proceeds as usual under the GLM (Kutner et al., 2005). This is implemented in FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) under `OLS` while in SPM (Wellcome Trust Centre for Neuroimaging U.C.L, 2010) this is the standard implementation. In AFNI (Cox, 1996) this is implemented under `3dttest++` (see also Chen, Saad, Nath, Beauchamp, & Cox, 2012).

WLS: allowing for heteroscedasticity The WLS approach, or more generally the Generalized Least Squares (GLS) approach, explicitly models the two components of the variance-covariance of $\boldsymbol{\eta^*}$ in (4.6):

$$\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}} = \begin{bmatrix} \sigma_1^2 + \sigma_\eta^2 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma_M^2 + \sigma_\eta^2 \end{bmatrix} \tag{4.8}$$

More specifically, a weighting matrix W is constructed such that more variable estimates b_m are down-weighted in the estimation of $\boldsymbol{\gamma}$. In the special case where the design matrix $\boldsymbol{X_m}$ only consists of a column of 1's, the closed form expression for the estimator of $\boldsymbol{\gamma}$ equals (Mumford & Nichols, 2009)

$$\hat{\boldsymbol{\gamma}}_{\text{WLS}} = \sum_{m=M}^{M} \frac{b_i}{\sigma_m^2 + \sigma_\eta^2} \left(\sum_{m=1}^{M} \frac{1}{\sigma_m^2 + \sigma_\eta^2} \right)^{-1} \tag{4.9}$$

More generally, $\hat{\boldsymbol{\gamma}}_{\text{WLS}}$ equals:

$$\left(\boldsymbol{X_m^t} \hat{\boldsymbol{W}} \boldsymbol{X_m} \right)^{-1} \boldsymbol{X_m^t} \hat{\boldsymbol{W}}^{-1} \boldsymbol{b} \tag{4.10}$$

with $\boldsymbol{W}$ the weighting matrix:

$$\boldsymbol{W} = \begin{bmatrix} \left(\sigma_1^2 + \sigma_\eta^2\right) & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \left(\sigma_M^2 + \sigma_\eta^2\right) \end{bmatrix} \tag{4.11}$$

Inference for the variance components is more complex since no closed form solutions exist. Several (restricted) maximal likelihood approaches have been suggested in the literature (see e.g. Chen et al., 2012). In practice, the within-subject variance is often set to the first-level variance estimates (Mumford & Nichols, 2009, also in the FSL software package).

In FSL this is implemented under `Flame1` while in AFNI this is implemented under `3dMEMA` (see also Chen, Saad, Britton, Pine, & Cox, 2013).

4.2.2 Dealing with the Multiple Testing Problem

It is well-known that the mass-univariate approach in which V ($V >$ 100.000) voxels are tested simultaneously is faced with huge multiple test-

ing problem, even at the second level. Indeed, if 100.000 tests for which H_0 is true are conducted simultaneously, each at a significance level of $\alpha = 0.05$, then, by chance alone, 5000 voxels will be declared active. Hence, the number of false positives (FP, see Table 4.1) becomes unacceptably high. While the interest lies in minimizing both the number of FPs and false negatives (FNs), multiple testing procedures aim to control FP rates (type I error rates).

		Decision	
		Conclude H_0	Conclude H_1
Voxel	Active	False Negative (FN)	*True Positive* (TP)
	Inactive	*True Negative* (TN)	False Positive (FP)

Table 4.1 Table of events for Null Hypothesis Significance Testing (NHST) in which evidence against a null hypothesis H_0 is evaluated in the direction of an alternative hypothesis H_1.

Familywise Error Rate (FWE)

The FWE is the probability that at least one FP occurs among all tests performed (see e.g. Nichols & Hayasaka, 2003). In order to control this error rate, one needs the null distribution of the maximum statistic over the V test statistics: $max(T_v)$. Indeed, assuming that the global null (i.e. the null hypothesis holds for all voxels) holds, we have that

$$P\left(FP > 0 \mid \text{global } H_0\right) = P\left(\bigcup_{v=1}^{V} T_v > u \mid \text{global } H_0\right) \tag{4.12}$$

$$= P\left(\max\left(T_v\right) > u \mid \text{global } H_0\right) \tag{4.13}$$

Hence, when u is chosen such that this probability is lower or equal to α, the FWE is controlled at level α. In fMRI data analysis, the most commonly used approach to control the FWE is based on Random Field Theory (RFT, see e.g. Brett, Penny, & Kiebel, 2007). Relying on parametric assumptions, RFT allows a closed form approximation of the upper tail of the null distribution of the maximum statistic. Alternatively, non-parametric methods for inference such as permutation-based testing, may be used. In the latter case. This will be discussed more extensively in the section 4.2.3.

Note that the expressions in Equations (4.12) and (4.13) imply weak control of the FWE as control is only guaranteed under the assumption that the null is true for all voxels. (Nichols & Hayasaka, 2003, section 2.3) argue that in imaging this weak control of FWE also entails strong control; i.e. control for any subset of null voxels. This is essential to localize individual significant voxels.

Further note that the classical Bonferroni correction, in which the observed p-value is multiplied with the number of tests and compared with to α, can also be used to control the FWE. The underlying assumption of independence when using the Bonferroni correction implies very conservative results in the fMRI context however, and makes the Bonferroni correction relatively useless. While corrections for dependence exist, these are seldomly used in the analysis of neuroimaging data (Nichols & Hayasaka, 2003).

False Discovery Rate (FDR)

FWE is a very stringent error rate and controlling it leads to conservative corrections. Given that one is willing to accept more FPs, provided that this number is small relative to the total number of selected voxels, one can rely on a different error measure, the False Discovery Rate (FDR). The FDR equals $E(Q)$ with

$$Q = \begin{cases} \dfrac{\#FP}{\#\text{selected voxels}} = \dfrac{\#FP}{\#FP + \#TP} & \text{if \# selected voxels} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

Genovese et al. (2002) introduced a procedure to control the FDR in neuroimaging. Using the procedure of Benjamini & Hochberg (1995), the FDR is considered at level q in the sense that

$$E(Q) \leq \frac{\#FP + \#TN}{V} q \leq q \tag{4.15}$$

The algorithm is as follows (Genovese et al., 2002):

1. Select a level q
2. Order all V original p-values from smallest to largest. With ℓ_v representing the v^{th} smallest p-value, i.e. $p_{\ell_v} = p_{(v)}$, the ordered p-values are as follows:
   $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(V)}$

3. Define r such that it is the largest v for which $p_{(v)} \leq \frac{v}{V}q$ holds.

4. Declare all voxels $\ell_1 \dots \ell_r$ active

Genovese et al. (2002) argue that this procedure controls the FDR under the assumption of *positive dependence*, i.e. noise is Gaussian with non-negative correlation. This assumption is reasonable given that smoothing images imposes increased dependency between neighboring voxels (and thus tests).

Uncorrected threshold with minimum cluster size

Based on simulation studies, Lieberman & Cunningham (2009) proposed a more ad-hoc two-step procedure that aims for a better balance between FP and FN. In the first step, the test image is thresholded at u, corresponding to an uncorrected α of e.g. 0.005. In the second step, only those voxels belonging to a cluster with minimal cluster size of 10 are selected.

Relation with cluster-based significance testing It should be noted that the method of an uncorrected threshold with a minimum cluster size shows superficial resemblances with cluster-based significance testing procedures. Cluster-based significance testing is a popular method to detect activation (Woo et al., 2014). It is however fundamentally different in nature from the procedures described above. Indeed, it uses topological features rather than purely voxel-based characteristics and therefore relies on different assumptions.

As suggested by the reviewers, we added this method to our comparison in the simulations for completeness (see Section 4.3). More specifically, we added the cluster size (S) based significance testing with FWE-corrected and FDR-corrected p-values. This corresponds to the two-step procedure but the minimum cluster size S is obtained based on cluster significance instead of fixing it at 10. Similarly to the two-step procedure, a first threshold α is chosen and only clusters that are sufficiently large are retained as significant. Without going into technical details for both permutation-based and parametrical inference (wich can be found in e.g. Woo et al., 2014; Friston, Holmes, Poline, Price, & Frith, 1996; Hayasaka & Nichols, 2003), this procedure determines the significance of a cluster in order to obtain the minimum cluster size S. More specifically, in a first step, after having set a sufficiently high fixed first threshold (e.g. $\alpha = 0.001$), clusters are determined by a cluster-forming algorithm. In

a second step, for each of these supra-threshold clusters, the probability to observe a cluster of size S under the null hypothesis of no activation can be determined. These cluster p-values can be corrected to control either the FWE (further referred to as cluster-FWE) or the FDR (further referred to as cluster-FDR) at cluster level.

In the two-step procedure with a fixed cluster size of 10, the first threshold α can be varied (empirically). For cluster-based inference on the other hand, it is important to note that the null distribution of cluster sizes relies on the assumption that the first (cluster-forming) threshold remains fixed at a stringent α-level, typically of $\alpha = 0.001$. This implies that in the simulations, it is the minimum cluster size S that is varied empirically for the cluster-based approach (by imposing different statistical thresholds for cluster sizes through varying the FWE or FDR) and not the cluster-forming threshold α.

4.2.3 Inference

Parametric inference

If one is willing to make distributional assumptions for the test statistic of interest, one can easily derive the thresholds for inferential decision making. We first discuss such parametric inference for the FWE, and next for the FDR and the two-step approach.

For the FWE correction, one can rely on Random Field Theory (RFT) to derive the null distribution of $max(T_v)$. Using two essential approximations from *Gaussian* Random Field Theory (which we will not discuss in full detail here, more details can be found elsewhere e.g: Brett et al., 2007; Nichols & Hayasaka, 2003), we have that:

$$FWE = P\left(max(T_v) > u \mid \text{global } H_0\right) \tag{4.16}$$

$$\approx P(\chi_u > 0) \tag{4.17}$$

$$\approx E(\chi_u) \tag{4.18}$$

In Expression (4.17), the FWE is approximated by the probability that the Euler Characteristic χ_u is larger than 0. χ_u basically counts the number of clusters under the null hypothesis, i.e. a collection of neighboring voxels for which $T_v > u$ holds. If the cluster-forming threshold u is set sufficiently high the probability to observe more than 1 cluster is neglected and one can approximate the FWE with Expression (4.18). The expected value of χ_u is estimated through a closed-form approximation that uses

information about the smoothness of the image of test statistic (Brett et al., 2007; Nichols & Hayasaka, 2003). The method not only takes into account the spatial character of the data through the smoothness, but also its computational efficiency is a major advantage (Nichols, 2012). It is challenging however to satisfy the main underlying assumptions needed for valid inference: i.e. normally distributed noise, sufficient smoothing and a sufficiently high threshold (see e.g. Worsley, Evans, Marrett, & Neelin, 1992; Brett et al., 2007).

For the FDR corrected inference and the two-step procedure, uncorrected p-values that are based on the usual t distributions of the test statistics which rely on normally distributed noise, as obtained from the OLS and WLS approach, can simply be used.

Permutation-based inference

Although some tools exist to verify the distributional assumptions underlying the test statistic (e.g. Luo & Nichols, 2003), there is no widespread tradition to check those assumptions in fMRI data analysis (Monti, 2011). The parametric null distributions indeed often rely on strong assumptions, which are seldom entirely fulfilled (A. P. Holmes et al., 1996). Therefore one could alternatively use non-parametric approaches such as bootstrap (e.g. Friman & Westin, 2005; Bellec, Rosa-Neto, Lyttelton, Benali, & Evans, 2010; Roels, Moerkerke, & Loeys, 2015) and permutation procedures (e.g. Nichols & Holmes, 2002; Thirion et al., 2007; Adolf et al., 2014). Using resampling techniques, the permutation approach for example guarantees (asymptotically) valid inference at nominal levels by creating a null distribution conditional on the observed data, but that advantage comes at the cost of increased computational effort.

Focusing on second-level analysis, and the scenario where one simply wants to test for activation over all individuals (i.e., the design matrix $\boldsymbol{X_M}$ is a vector of 1's), permutation-based testing proceeds as follows:

1. Define P: the number of permutations, the higher P, the higher the precision of the empirical null distribution. However, the computational burden also increases with increasing P.

2. Compute for each voxel v the test statistic in the original sample: T_{v0} for each voxel

3. Create P new samples by randomly flipping the sign of some of the elements in $\boldsymbol{X_M}$, i.e. for randomly chosen individuals the 1 is

changed into -1 (A. P. Holmes et al., 1996). [1]

4. For each of the P (with $p = 1, \ldots, P$) samples compute the test statistic T_{vp}

5. The permutation null distribution for voxel v is then defined as the empirical distribution of the T_{vp}'s. Clearly, the smaller the number of permutations P is, the more discrete the null distribution will be.

Within a mass-univariate approach, empirical p-values are obtained per voxel using $P\left(T_{pv} \geq T_{v0}\right)$, the probability to observe a test statistic in the permutation null distribution that is at least as large as the test statistic observed in the sample at hand. The FDR correction and the two-step procedure are performed on these p-values.

For the FWE correction, permutation-based inference proceeds via the empirical sampling of the maximum statistic over all voxels to obtain the null distribution of the maximum statistic. This implies that in step 4 the maximum over the test statistic of all voxels is calculated: $T_p = max(T_{pv})$ with $(v = 1, \ldots, V)$.

4.3 Simulations

4.3.1 Data Generation

For every subject $(m = 1, \ldots, 15)$ and for every voxel in a 3-dimensional space $(45 \times\ 45 \times\ 45)$, we generate a time series $\boldsymbol{y}$ for the signal on the first level using the following model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{d} + \boldsymbol{\epsilon}, \tag{4.19}$$

with $\boldsymbol{\beta} = [\beta_0, \beta_1]^t$ and with $\boldsymbol{X}$ the design matrix, consisting of a column for the intercept and a column describing the expected signal under a simple block design. $\boldsymbol{Z}$ is identical to $\boldsymbol{X}$, and $\boldsymbol{d}$ contains a random intercept d_0 and random slope d_1. The random intercept variance was set to zero, while a random slope d_1 is drawn from $N(0, \sigma^2_{d_1})$ for every subject to allow for heterogeneous effects of $\boldsymbol{X}$ on $\boldsymbol{y}$ between subjects. For every subject, voxel and time point, ϵ is drawn from $N(0, \sigma^2_m)$. In the simulation study no temporal correlation was induced as this unnecessarily might influence

[1] If the individuals belong to different groups, or the study design is more complex, more appropriate schemes can be found in e.g. Winkler, Ridgway, Webster, Smith, & Nichols (2014).

our variance estimates and consequent inference (see e.g. Lenoski, Baxter, Karam, Maisog, & Debbins, 2008, for an investigation of the impact of modeling the temporal autocorrelation in fMRI). We further define a signal to noise ratio (SNR) as the maximum amplitude ($\boldsymbol{x}\beta_1$) divided by σ_{d_1} and focus on a simple contrast $\boldsymbol{c\beta}$ with $\boldsymbol{c} = [0, 1]$

The between-subjects standard deviation, σ_{d_1} was set such that the SNR=1 (low signal strength) or the SNR=2.5. The variance σ_m^2 is either constant or varying over the M subjects. To ensure comparability between both scenarios in terms of the average total amount of variability, the variance σ_m^2 under the constant scenario is set to the average of all values under the varying scenario.

We use the neuRosim R package (Welvaert & Rosseel, 2012) and a canonical HRF to set up the first level activation (Henson & Friston, 2007) in Equation (4.19). In total there are 1934 active voxels, distributed over two clusters, and 89191 inactive voxels in a $45 \times 45 \times 45$ volume ($\pm 2.5\%$ of the voxels). The noise images that were added to the activation image, were minimally smoothed in order to comply with the basic assumptions for RFT (Monti, 2011; Lindquist, 2008; Brett et al., 2007).

In total, 1000 simulations are performed for all 4 data generating mechanisms (2 SNR, and constant versus varying σ_m^2).

4.3.2 Analysis and Evaluation Details

Analysis

We focus on the OLS and WLS approach to combine the individual evidence from the M subjects. FSL (version 5.0.7, Jenkinson et al., 2012), one of the most frequently used software packages to analyze fMRI data (Carp, 2012), has both methods implemented. First, the estimates $\boldsymbol{c}\hat{\boldsymbol{\beta}}$ (see Equation (4.1)) are obtained and next used for the second-level analysis. In the WLS approach, for every subject m σ_m^2 is estimated (see Equation (4.6)) and then used to weight the evidence per subject as outlined in Equation (4.11). For the parametrical inference in the OLS case, inference is based on the t distribution with $M - 1$ degrees of freedom. The WLS method uses an intrinsic Bayesian procedure that takes into account both the subject-specific variability and the variability on the estimation of $\boldsymbol{c\beta}$. Further inference proceeds via a back-transformation of the posterior probability $P(\boldsymbol{c\gamma} > 0|\boldsymbol{b})$ (see Equation (4.3), and Mumford & Nichols, 2006) to a Z-map.

For both the OLS and the WLS we use the permutation technique

based on *sign-flipping*, see Section 4.2.3. The command line tool `randomise` allows for permutation-based on the OLS method. For the WLS approach we followed the same protocol, but via an in-house `R` script with the test statistic as in Equation (4.9). The permutation null distributions are based on 5000 permutations. On a standard laptop computer the computational time for the OLS permutation was less than 10 minutes compared to over about 40 minutes for the WLS permutation. We note that compared to the FSL implementation our in-house script was not fully optimized to speed up computational time.

Evaluation

The performance of the different combination of techniques is evaluated based on the Receiving Operating Characteristics (ROC) curves. The ROC curves show the true positives (TP) rate in function of the false positives (FP) rate, with the FPs defined as voxels that are declared active but not in the true activation region and the TPs as the voxels that are declared active and in the true activation region.

ROC-curves provide a means to investigate the balance between the FP and TP rate, however, bias may be introduced for imbalanced data. As in fMRI, there are typically more true inactive than true active voxels, we also provide the Matthews correlation coefficient (Matthews, 1975). This measure takes into account the four cells as displayed in Table 1 and is therefore a more comprehensive measure for the quality of a test criterion, even for imbalanced data (see e.g. Vihinen, 2012, for an application in the genetical context). The Matthews correlation coefficient (MCC) is calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.20}$$

Values close to 1 indicate more correct decisions, values close to 0 indicate random decisions, and values close to -1 indicate more incorrect decisions.

Furthermore we study stability through the variation on the number of correctly selected voxels. Stable methods are methods that do not induce much variability on the number of selected voxels. At last, from the above, it should be clear that all measures are defined in voxel-based way.

4.3.3 Results

In Figure 4.1 we present the ROC curves under each of the four data generating mechanisms (low versus high SNR in left versus right panel, equal versus unequal σ_m^2 in the upper versus lower panel). In total 12 ROC curves are presented, one for each of the $2 \times 2 \times 3$ combinations of selection procedures (OLS versus WLS, parametric versus non-parametric inference, FWE versus FDR versus 2-step procedure). We summarize the most important findings below.

First, we find that under all scenarios the two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10 (further denoted as BCL) has a better trade-off between FP and TP than the FWE-control or FDR-control.

Second, both under high and low signal strength, the ROC of the permutation-based method and the parametric inference have very similar shapes at almost the same height when focusing on the OLS approach. When considering the WLS approach, one finds that the ROC curves are substantially higher with permutation-based inference than with the parametric inference under both SNR's (regardless of the type of control).

Third, in almost all panels of Figure 4.1 we find a good performance of the WLS versus the OLS method under the parametric approach, regardless of the type of multiplicity control. When permutation-based inference is used a similar performance of OLS and WLS is observed when the SNR is low, but it the WLS seems to perform worse than OLS when the SNR is high. It should be noted that this is due to the discreteness of the permutation-based inference, which is mostly apparent when the signal is strong.

In Figure 4.2, the MCC is depicted for respectively a low and high signal strength with respect to the total number of selected voxels (FP+FN). While the findings based on the pattern of the ROC-curve are mostly confirmed in these figures, the differences under high SNR are somewhat less pronounced. This may indicate that under high SNR the decisions diverge less than when the SNR is lower for a same number of selected voxels.

Figure 4.3 shows the proportion of correctly selected voxels on the X-axis and its corresponding standard deviation on the Y-axis. For all 4 data-generating mechanisms, we find that the FDR correction for multiple testing results in more variability than the other two procedures that correct for multiple testing. We also find that the FWE correction results in slightly more variable results than the BCL based corrections. Furthermore, this pattern is not altered by the choice for permutation-based

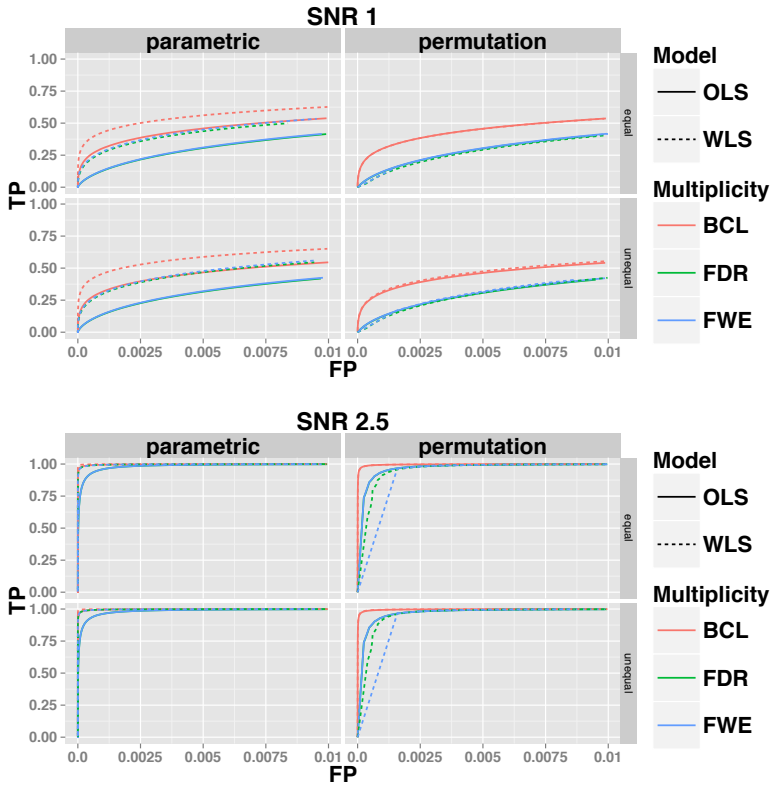


Figure 4.1 ROC for the low signal strength (SNR=1) and for the high signal strength (SNR=2.5); for differences in the subject-specific variability (*unequal*) or identical subject-specific variability (*equal*); for permutation-based inference and for parametric inference. FWE: family-wise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

inference or parametric inference. One exception is however observed. Indeed, we find that for the WLS procedure, under the high SNR, the BCL procedure becomes more variable than the FWE procedure. We attribute this, again, to the discreteness of the permutation method and the high signal present in this simulation.

Figure 4.4 depicts the comparison between the BCL procedure and the

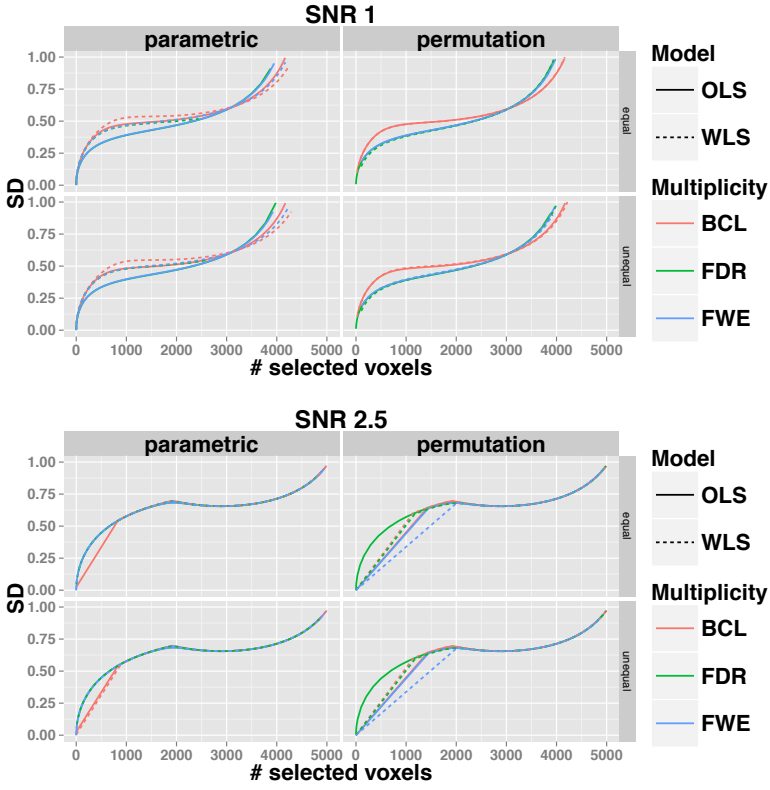


Figure 4.2 Matthews Correlation Coefficient (MCC) for the low signal strength (SNR=1) and for the high signal strength (SNR=2.5); for differences in the subject-specific variability (*unequal*) or identical subject-specific variability (*equal*); for permutation-based inference and for parametric inference. FWE: family-wise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

pure cluster-size based inference in the ROC-curve in the simulations with no between-subject differences in the residual variability. The results for the case *with* differences in the within-subject variability, and the results for the Stability plots and the MCC are presented in appendix B. We note that due to the first fixed threshold in pure cluster-based testing, the maximum number of selected voxels is limited. For the ROC-curves and for the stability we find discrete patterns. These are a logical consequence

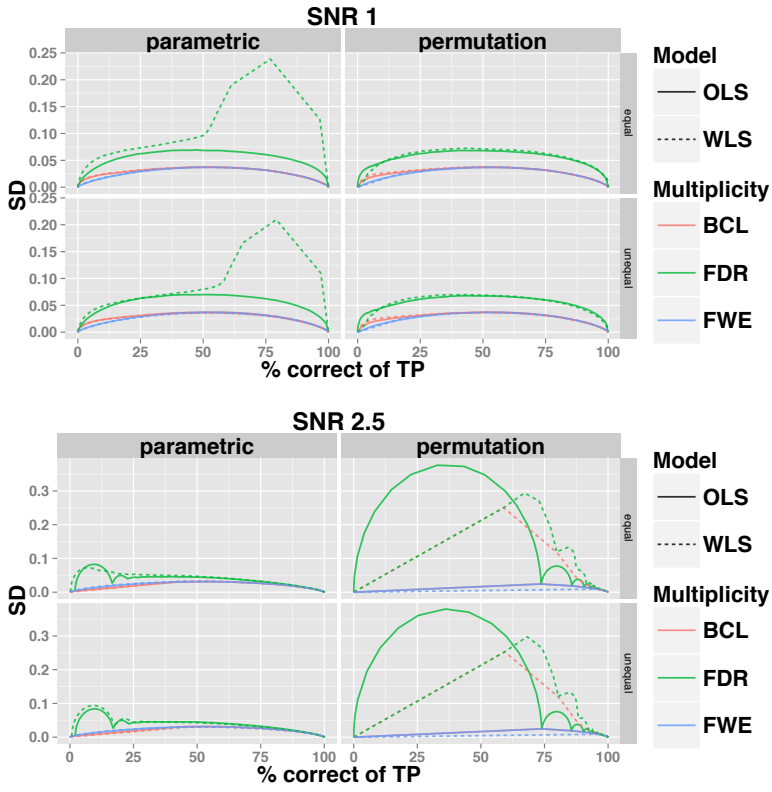


Figure 4.3 Stability plot for the number of correctly selected voxel in the simulation with low signal strength (SNR = 1) and for the high signal strength (SNR=2.5); for differences in subject-specific variability (unequal) or identical subject-specific variability (equal); and for permutation-based inference and for parametric inference. FWE:family-wise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

of our simulation setup, in which two relatively large clusters are set active. Based on the ROC-curve we find a good trade-off between FP and TP for the cluster-based inference when the SNR is high, but not when the SNR is low. For the stability, it is hard to draw conclusions based on the observed results due to the above mentioned limitations.

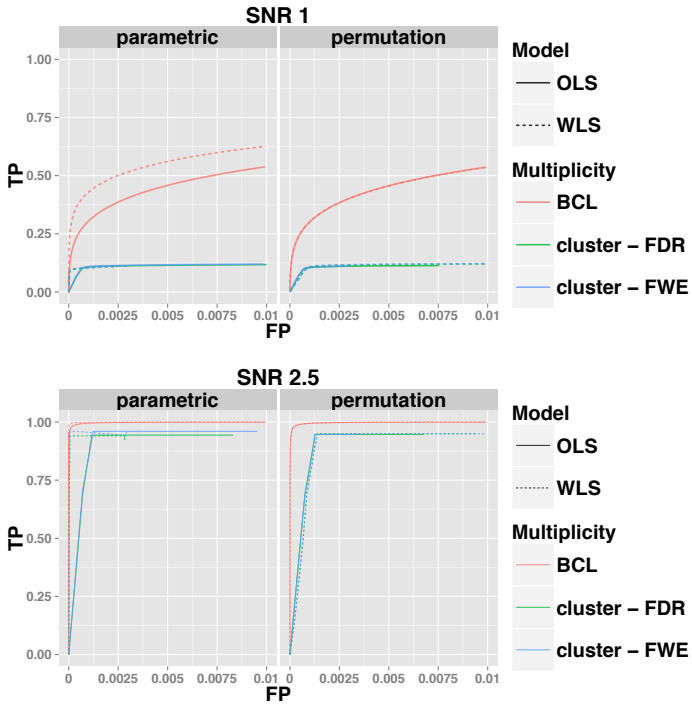


Figure 4.4 ROC for the low signal strength (SNR=1) and for the high signal strength (SNR=2.5) with identical subject-specific variability; for permutation-based inference and for parametric inference for cluster-based inference with $\alpha = 0.001$: cluster - FWE: family-wise error correction based on cluster-size inference, cluster - FDR: False Discovery Rate correction based on cluster-size inference, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

Finally note that under the lowest signal strength, we find a peak in the variability for the WLS approach in combination with the FDR correction. Further inspection of the p-values for the WLS approach reveals that this is due to more discreteness in the highest p-values compared to the OLS procedure (Figure 4.5).

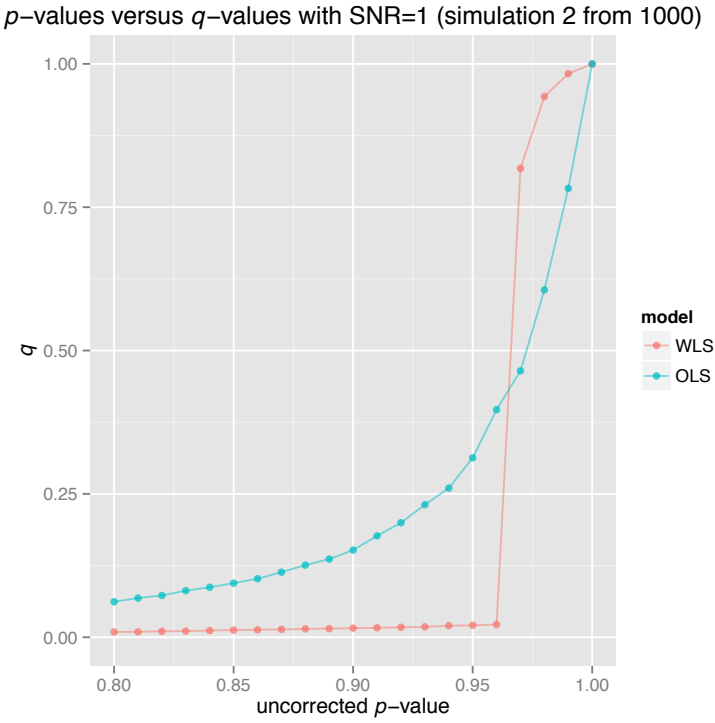


Figure 4.5 Uncorrected p-values for the OLS and the WLS procedure, with their corresponding FDR corrected q-values based on one specific simulation under SNR=1 with equal variance among subjects.

4.4 Real Data Example

4.4.1 Human Connectome Project Dataset

To check the findings from the simulation study on real data, we use data from the Human Connectome Project (hCP, Van Essen et al., 2012). Those data are analyzed on the first level, using a standard protocol that is described elsewhere (Glasser et al., 2013). To mimic a typical fMRI study with about 15 subjects, we select the first 15 subjects[2] from the HCP dataset with a focus on contrast 4, which entails the difference between a mathematical task and a story-telling task.

[2]Subject identifiers can be found in Appendix A.

4.4.2 Stability of the Selected Voxels

For the HCP data, we determine the stability of the different proposed methods by bootstrapping subjects from the original sample, i.e. drawing subjects with replacement from the original sample. In total, 100 bootstrap samples are taken. The number of active voxels at level 2 is determined in each these bootstrapped datasets, using one of 12 the aforementioned combinations for inference at the second level. The stability on the number of selected voxels over bootstrap samples is further assessed by considering the *re-selection rate* of a specific voxel, which is the proportion of bootstrap samples in which that voxel is declared active.

4.4.3 Results

In Figure 4.6, we find the same pattern as in the simulations when using parametric inference: i.e. the FDR based correction for multiple testing results in more variability on the number of selected voxels. Also, we find that the FWE and the BCL correction result in similar variability. This finding holds both for the WLS and the OLS approach. In contrast to the simulation study, we find however that the WLS approach is always less variable than the OLS approach for a given type of multiplicity control.

For the permutation-based inference we find that when the number of selected voxels is relatively low (less than $\pm$ 5% of the $\pm$ 200.000 voxels) the FDR correction with the OLS is far more variable than all other combinations. We note again that the WLS suffers from the discreteness of p-values in the permutation-based inference when the FDR correction is used. Due to this discreteness, several small original p-values are converted to only one corrected q-value, causing the straight line from the origin to the first point. For the two-step procedure, there is a similar artifact when using WLS. This can be attributed to the fact that the lower p-values do not occur in clusters larger than 10, until these reach a certain threshold that results in a huge amount of activation. If more than 5% of the voxels are selected, the results are more variable if one uses the FWE correction for multiple testing, compared to the other methods.

Based on Figure 4.6, we next determine the thresholds for which 10.000 voxels are selected on average over the 100 bootstrap samples. These threshold are then used to determine the re-selection rate of the each specific voxel over the 100 bootstrap samples. Figure 4.7 depicts the histograms of the re-selection rates that are larger than 50%. The header of each histogram shows the percentage of voxels that are selected in more

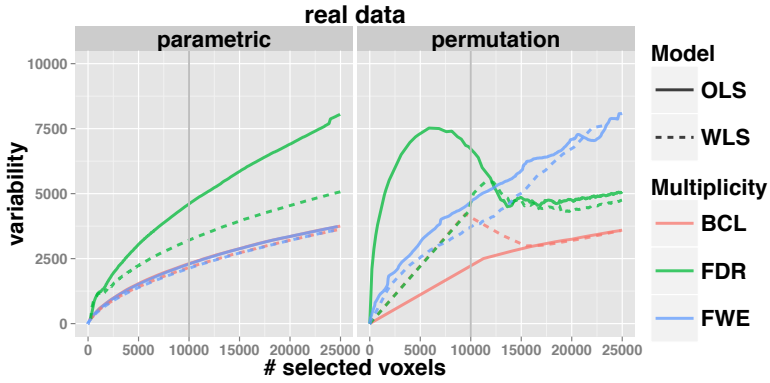


Figure 4.6 Stability plot the number of selected voxels for $n = 15$ of the HPC dataset for permutation-based inference and for parametric inference. FWE:familywise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

than 90% of the samples.

From Figure 4.7 we find the highest re-selection rates when using the FWE or BCL multiplicity control in the parametric inference framework (i.e., the 6 upper panel histograms). In the permutation-based inference framework (i.e., the 6 lower panel histograms), we find that the FDR achieves higher re-selection rates than the FWE if the OLS approach is used, but the highest re-selection rates are found with the BCL multiplicity control both with the OLS and the WLS approach.

To take into account the localization of voxels that are frequently re-selected, we also constructed brain images in Figure 4.8, where we identified all voxels that have a re-selection rate of at least 75%. Although we acknowledge that the slice depicted is only exemplary, the above described trends are clearly confirmed.

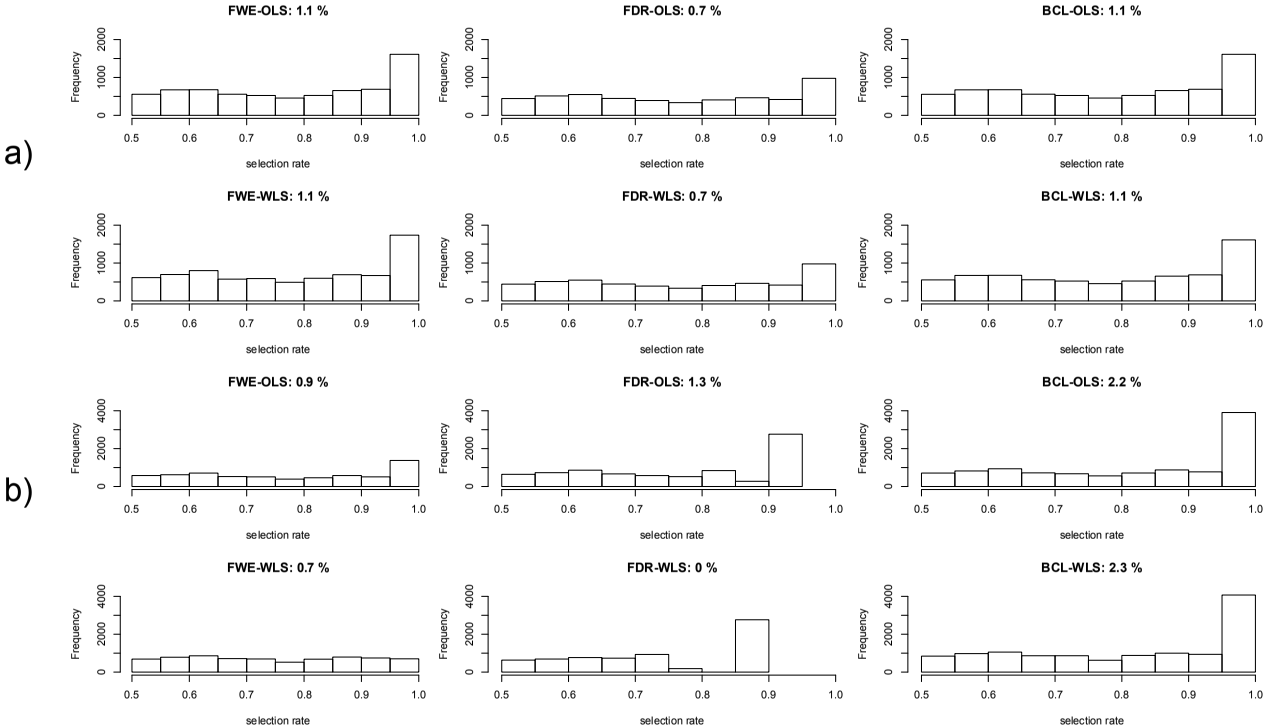
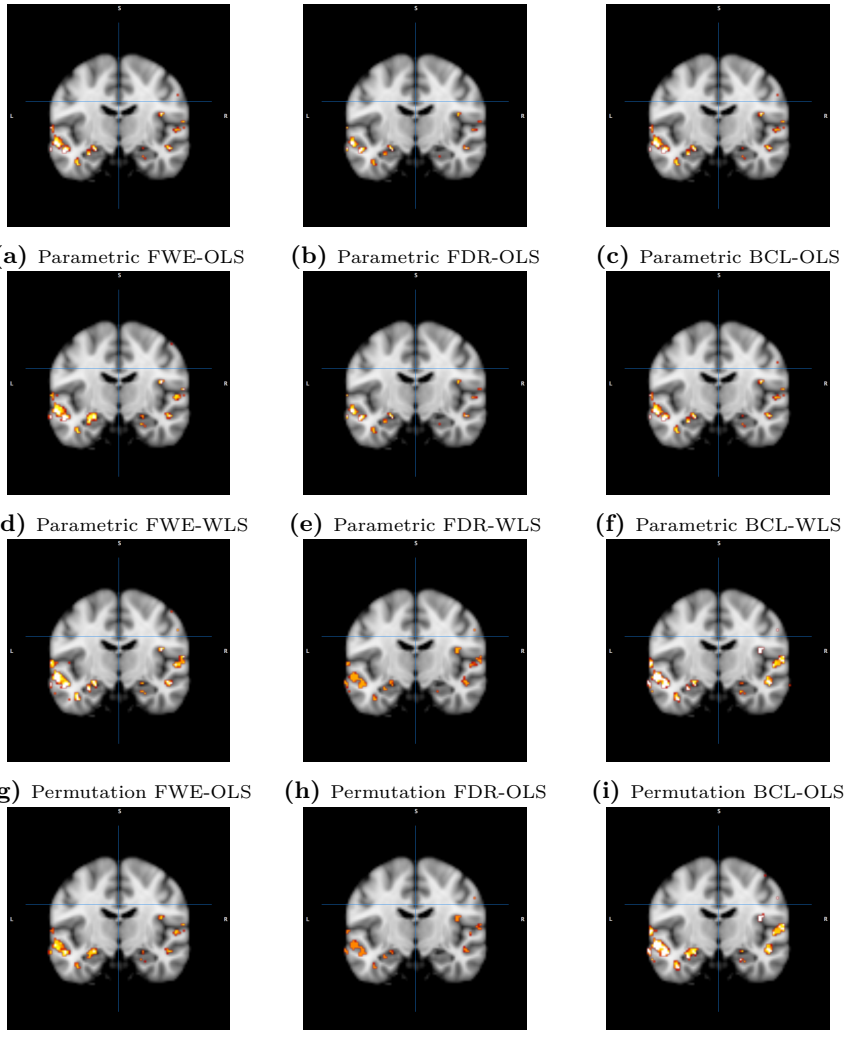


Figure 4.7 Plot with the re-selection rates of the voxels that are larger than 0.5 over 100 bootstrap samples for real data for parametric inference (a) and for permutation-based inference (b). FWE:familywise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach. The indicated percentage denotes the number of voxels that are declared active in more than 90 % of the bootstrap cases.



(a) Parametric FWE-OLS **(b)** Parametric FDR-OLS **(c)** Parametric BCL-OLS

(d) Parametric FWE-WLS **(e)** Parametric FDR-WLS **(f)** Parametric BCL-WLS

(g) Permutation FWE-OLS **(h)** Permutation FDR-OLS **(i)** Permutation BCL-OLS

(j) Permutation FWE-WLS **(k)** Permutation FDR-WLS **(l)** Permutation BCL-WLS

Figure 4.8 Plot with the re-selection rates that are larger than 0.75 for the HPC data for parametric inference (the top two rows) and for permutation- based inference (the bottom two rows). FWE:familywise error correction, FDR: False Discovery Rate correction, BCL: Bonferroni-like first threshold and minimal cluster size. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach. The average number of activated voxels was kept constant for all cases. Red/orange: closer to 0.75; white:closer to 1.

4.4.4 Test-retest Correspondence

As suggested by one of the reviewers, stable methods should reflect more similar results using different real samples. To study this, we used an additional run for each of the 15 subjects in the HCP data. We exemplary demonstrate this test-retest similarity for the parametrical analysis. We matched the number of selected voxels per image in the FWE/FDR method by the respective numbers that are found using the two-step BCL procedure. Indeed, when selecting the N voxels with the N smallest p-values, the FWE and FDR method results are identical. This matching on the number of selected voxels is motivated by the simulation findings that larger a number of selected voxels results in a higher MCC. In a test-retest setting, the MCC coincides with the correlation between two binary images (selected / non-selected voxels). In figure 4.9 we see that indeed the BCL outperforms the FDR/FWE and that the WLS outperforms the OLS. We note however that this methods has a major drawback as it does not allow to calculate the variability on these numbers and it requires a second sample.

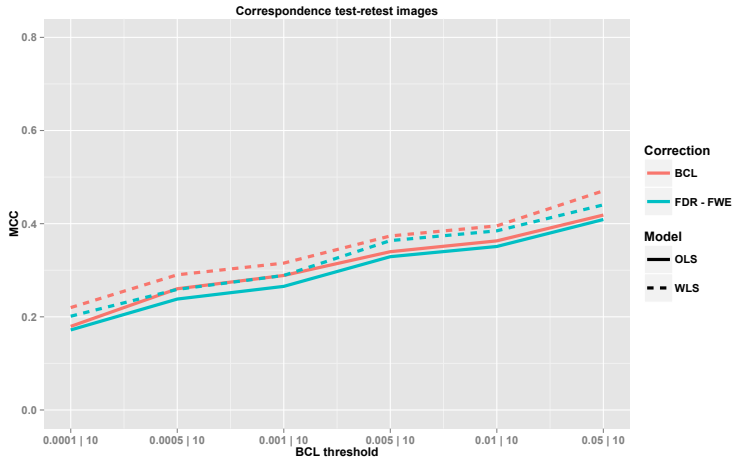


Figure 4.9 Test-retest correspondence measured trough the correspondence between two binary images (selected/non-selected voxels). Each BCL threshold corresponds to a specific number of selected voxels which may vary between images but not between methods.

4.5 Discussion

In this study we investigated both the balance between true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and data analytical stability of methodological choices in the second-level analysis of fMRI data. Following the traditional evaluation of techniques in the fMRI literature, we first focused on the balance between FP and TP, using ROC-curves and on the Matthews correlation coefficient (MCC), a measure that takes all possible decisions into account. Aiming for more reproducible brain imaging research, we believe however that data analytical stability is also an important criterion that offers an additional unique perspective on the behavior of methods. While studies using the criterion of data analytical stability are sparse and mostly focused on the first-level inferential decisions (e.g. Durnez, Roels, & Moerkerke, 2014; Roels, Bossier, et al., 2015, for respectively a focus on mass univariate inference and topological inference), this study filled this gap through considering data analytical stability of different methods at the second-level analysis. Unlike the NPAIRS framework (Strother et al., 2002, 2004) that allows to explore overall stability, we furthermore focused on the *selected* voxels, obtained via thresholded images, when assessing the data analytical stability.

More specifically, we assessed in this paper the impact of three different choices that the researcher has to make when analyzing fMRI data at the second level: (1) Should one use a WLS- or an OLS-approach, (2) Should one rely on parametric assumptions for the test statistic or rely on a non-parametric framework, such as permutation-based inference, and (3) Which type of control should one use to limit the multiplicity issue. The impact of these choices was assessed from the ROC-curves, MCC and the data analytical stability perspective.

For the balance in the decision context, based on the ROC-curves and the MCC, results were pretty clear when parametric inference is used. Regardless of the choice of the multiple testing correction, we found that the WLS-method yields a better the balance between FP and TP than when the OLS-method is used. While the MCCs confirmed most of the results based on the ROC-curves, they revealed the fact that differences are more obvious when the SNR was low. Under the high signal strength, the balance in the decision context did not diverge remarkably between methods. This findings on the balance between FP and TP are in line with Mumford & Nichols (2009), although the magnitude of the difference between

WLS and OLS was more pronounced, based on the ROC-curves, in our simulation study. When permutation-based inference is used, there were barely any differences between OLS and WLS. We found however that there were some effects of discreteness when permutation-based inference was used in combination with WLS. In the simulation settings this was associated with spiky patterns under a high SNR due to substantial jumps in the number of voxels that are selected. But also in the real data application, we found some evidence for discreteness with the WLS statistic when jumps in the activation occur. When comparing the parametric with the non-parametric approach, we found in contrast to Thirion et al. (2007) no evidence for a better performance of permutation-based inference. Note however that in all our simulation settings the basic assumptions of parametric inference were satisfied (Gaussian noise and sufficient smoothing). Upon inspection of the ROC-curves we also found in our simulation study that the two-step procedure, which ignores multiplicity first but requires a minimal cluster size next, outperforms the traditional FWE-control and FDR-control.

From a data analytical stability perspective, there were substantial differences between the three approaches we considered for multiple testing correction. In line with previous findings at the first level of analysis (Qiu et al., 2006; Durnez et al., 2014), FDR-based corrections for multiple testing resulted in more variable selections. Both in the simulation study and the real data application, we found that FWE based correction for multiple testing and a two-step procedure result in more stable results, as assessed by the variability on the number of selected voxels. This weaker performance of the FDR is observed, regardless of the WLS- versus OLS-approach, or the parametric versus non-parametric framework for inference. Interestingly, when we focused on the re-selection rate of a specific voxel in the data application, we also found superior performance of the two-step procedure. As noted by one of the reviewers, the increased stability for the FWE and two-step procedures relying on parametrical inference, might be attributed to the fact that these approaches exploit topological features of the data in contrast to the FDR.

While voxel-based inference is only one approach to control for multiple testing, several alternatives exist. Cluster-based inference (see e.g. Friston et al., 1996; Hayasaka & Nichols, 2003) is a very popular alternative that relies explicitly on topological features such as the cluster size and has been advocated because of the potential increase in power. However, Woo et al. (2014) showed that the commonly used two-step procedure for

cluster-based inference is non-robust when too liberal first thresholds are used at the voxel level, and that this results in unpractically large clusters when studies are sufficiently powered. This complicates the interpretation of the results as clusters could become as large as half of the hemisphere. In the same vein, Woo et al. (2014) and Nichols (2012) argue that the conceptual definition of a "significant cluster" is complicated by the fact that it is a randomly-sized collection of voxels of which one can only claim that at least some are significant. We concur with Nichols (2012) and Woo et al. (2014) that voxel-wise inference remains a useful alternative, and therefore opted for an extensive evaluation of commonly used voxel-based inference techniques.

The FP rates are evaluated only in a simulation study. While this might lack biological validity, this procedure allows to have strict control on the ground truth and consequent determination of TN and TP. With an exhaustive simulation study (2 SNR's and varying within-subject variability assumptions), we have covered some of the properties present in real data. Any simulation study comes naturally with the arbitrariness of these settings. However, compared to using real data to determine FP rates, simulation studies have the advantage to exclude unnecessary artifacts in the procedure to determine to the TP and the TN (see e.g. Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012, for differences in test errors based on the design) or its underlying assumptions.

Gathering all the above described evidence, we would recommend the brain imaging researcher to use WLS at the second level in combination with the two-step procedure, hereby relying on the parametric framework for inference. Note that throughout the paper, we have assumed that all images at the first level are correctly normalized such that individuals are perfectly co-registered. It should be stressed that further exploration of the robustness against violations of the parametric assumptions is warranted. However, the proposed strategy in this paper to assess data analytical stability of different methods on real data, could be used in any future application, and ultimately reveal the best choice from a data analytical stability perspective in practice. Such validation on real data may also yield further insight into the appropriateness of the rather ad-hoc but commonly used BCL-approach which lacks inferential justification.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercom-

puter Center), funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI.

Supplementary Material

Additional details HCP dataset

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

The list of subject identifiers used in this study can be found in Table S1.

100408	101915	103414
105115	106016	110411
111312	111716	113619
115320	117122	118730
118932	120111	122317

Table S1 The subject identifiers of the subjects included in the real data application. Subjects come from the 80 unrelated subjects dataset, release Q3 (Van Essen et al., 2012).

Figures

This section contains the additional figures in which the BCL procedures is compared with cluster-based inference procedures. For all of the following pictures the following abbreviations are used: 1) SNR = 1: low signal strength, SNR=2.5: high signal strength; 2) cluster - FWE: family-wise error correction based on cluster-size inference, cluster - FDR: False Discovery Rate correction based on cluster-size inference, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10; 3) OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach; 4) unequal: differences in the subject-specific variability, equal= identical subject-specific variability.

ROC-curves

In Figure S4.1 and Figure S4.2 the voxel-based ROC curves are depicted.

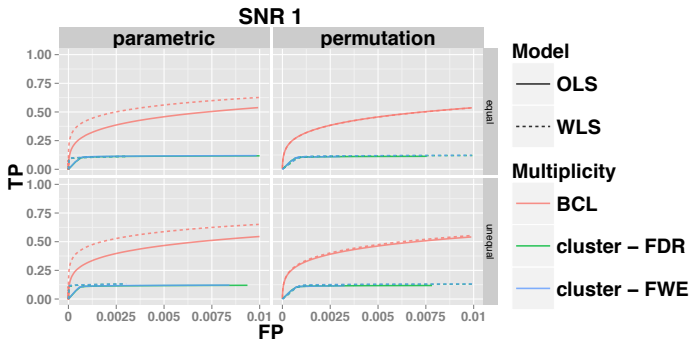


Figure S4.1 Receiver Operating Curve for a signal to noise ratio of 1 over the range $[0; 0.01]$.

Stability on the percentage of TP's

In Figure S4.3 and Figure S4.4 the voxel-based stability plots are depicted.

MCC

In Figure S4.5 and Figure S4.6 the voxel-based stability plots are depicted.

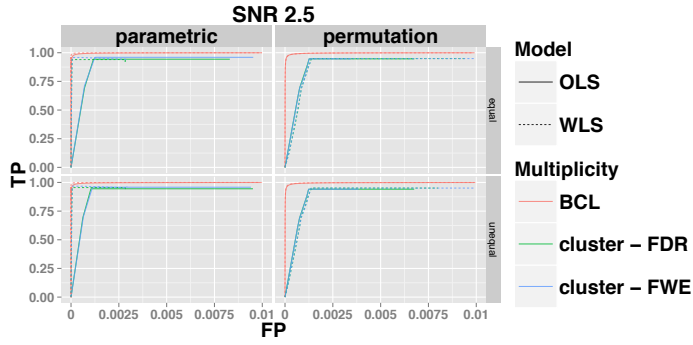


Figure S4.2 Receiver Operating Curve for a signal to noise ratio of 2.5 over the range $[0; 0.01]$.

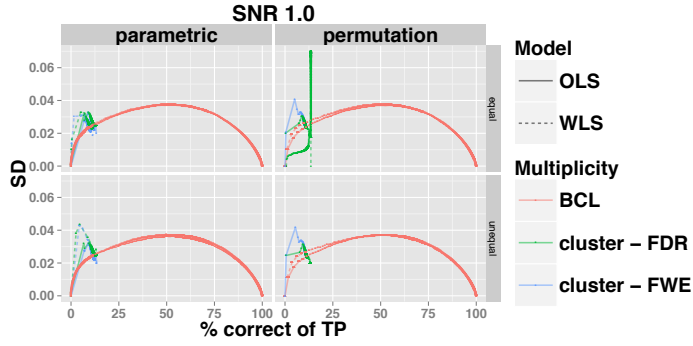


Figure S4.3 % of correctly activated voxels with their standard deviation for a signal to noise ratio of 1

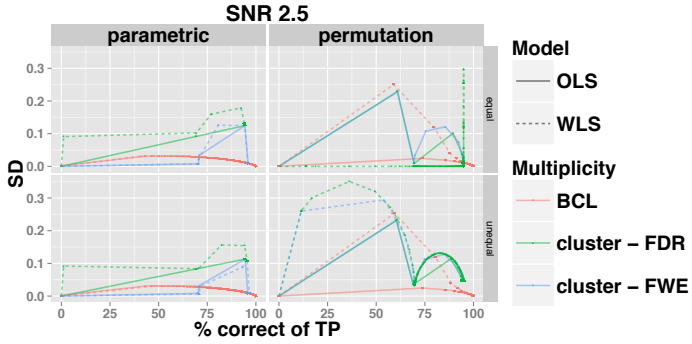


Figure S4.4 % of correctly activated voxels with their standard deviation for a signal to noise ratio of 2.5

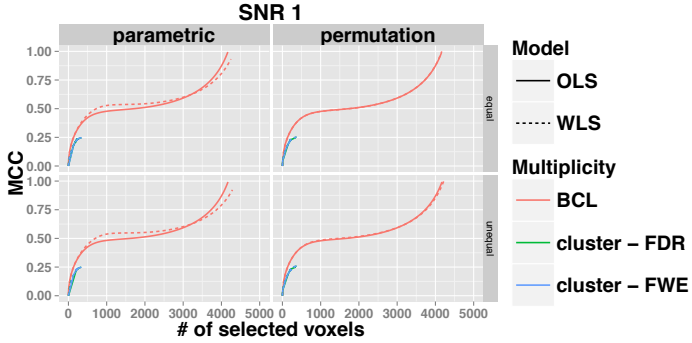


Figure S4.5 MCC for a signal to noise ratio of 1

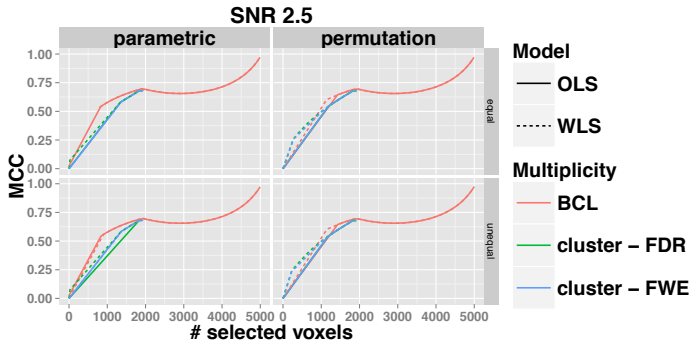


Figure S4.6 MCC for a signal to noise ratio of 2.5

References

Adolf, D., Weston, S., Baecke, S., Luchtmann, M., Bernarding, J., & Kropf, S. (2014). Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in neuroinformatics*, *8*, 72.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, *20*(2), 1052–63.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–1139.

Benjamini, Y., & Hochberg, Y. (1995). *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* (Vol. 57).

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, *4*(4), 417–22.

Brett, M., Penny, W., & Kiebel, S. (2007). Parametric procedures. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–90.

Chen, G., Saad, Z. S., Nath, A. R., Beauchamp, M. S., & Cox, R. W. (2012). FMRI group analysis combining effect estimates and their variances. *NeuroImage*, *60*(1), 747–65.

Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and biomedical research, an international journal*, *29*(3), 162–173.

Durnez, J., Roels, S., & Moerkerke, B. (2014). Multiple testing in fmri: a case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, *56*(4).

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Friman, O., & Westin, F.-J. (2005). Resampling fmri time series. *NeuroImage*, *25*, 859-867.

Friston, K. J. (2007). Functional integration. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 36). Elsevier Ltd./Academic Press.

Friston, K. J., Holmes, a., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, *4*(3 Pt 1), 223–235.

Friston, K. J., Holmes, A., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical Parametric Maps in Functional Imaging: A general Linear Approach. *Human Brain Mapping*, *2*, 189–210.

Genovese, C. R., Lazar, N. a., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*(4), 870–8.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., ... Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject fMRI test-retest reliability metrics and confounding factors. *NeuroImage*, *69*, 231–43.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Henson, R., & Friston, K. J. (2007). Convolution models for fmri. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 193-210). Elsevier Ltd./Academic Press.

Holmes, A., & Friston, K. (1998). Generalisability, random effects and population inference. In *NeuroImage* (Vol. 7, p. S754).

Holmes, A. P., Blair, R. C., Watson, J. D., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, *16*(1), 7–22.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Kiebel, S., & Holmes, A. P. (2007). The general linear model. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear models*. New York: McGraw-Hill Irwin.

Lenoski, B., Baxter, L. C., Karam, L. J., Maisog, J., & Debbins, J. (2008). On the performance of autocorrelation estimation algorithms for fmri analysis. *IEEE Journal of Selected Topics in Signal Processing*, *2*, 828-838.

Lieberman, M. D., & Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, *4*(4), 423–8.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464.

Luo, W.-L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, *19*, 1014-1032.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*, *405*(2), 442–451.

Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in human neuroscience*, *5*, 28.

Mumford, J. A., & Nichols, T. E. (2006). Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 42–51.

Mumford, J. A., & Nichols, T. E. (2009). Simple group fMRI modeling and inference. *NeuroImage*, *47*(4), 1469–75.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–5.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Poline, J.-B., & Brett, M. (2012). The general linear model and fMRI: Does love last forever? *NeuroImage*, *62*(2), 871–880.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, *7*.

Roels, S. P., Bossier, H., Loeys, T., & Moerkerke, B. (2015). Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis. *Journal of neuroscience methods*, *240*, 37-47.

Roels, S. P., Moerkerke, B., & Loeys, T. (2015). Bootstrapping fmri data: Dealing with misspecification. *Neuroinformatics*, *13*(3), 337-352.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., ... Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, *15*(4), 747–71.

Strother, S. C., La Conte, S., Kai Hansen, L., Anderson, J., Zhang, J., Pulapura, S., & Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage*, *23 Suppl 1*, S196–207.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, *35*(1), 105–20.

Vahdat, S., Maneshi, M., Grova, C., Gotman, J., & Milner, T. E. (2012). Shared and specific independent components analysis for between-group comparison. *Neural Computation*, *24*(11), 3052–3090.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, *13*(Suppl 4), S2.

Wellcome Trust Centre for Neuroimaging U.C.L. (2010). Spm 8 [Computer software manual]. http:// www.fil.ion.ucl.ac.uk/spm/.

Welvaert, M., & Rosseel, Y. (2012). How ignoring physiological noise can bias the conclusions from fMRI simulation results. *Journal of neuroscience methods*, *211*(1), 125–32.

Wilke, M. (2012). An iterative jackknife approach for assessing reliability and power of FMRI group analyses. *PloS one*, *7*(4), e35578.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–97.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, *91*, 412–9.

Worsley, K. J., Evans, a. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, *12*(6), 900–918.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*(1), 1–15.

5 Including Data Analytical Stability in Cluster-based Inference

Abstract A big challenge in the statistical analysis of functional Magnetic Resonance Imaging (fMRI) data is to account for simultaneously testing activation in over 100.000 volume units or voxels. A popular method that reduces the dimensionality of this test problem is cluster-based inference. We propose a new testing procedure that allows to control the family-wise error (FWE) rate at cluster level but improves cluster-based test decisions in two ways by (1) taking into account a measure for data analytical stability and (2) allowing voxel-based interpretation of results. For each voxel, we define the re-selection rate conditional on a given threshold and add this as a measure for stability into the selection process. Our procedure distinguishes between a liberal and conservative FWE controlling threshold. Clusters that survive the liberal but not the conservative criterion get selected if sufficient evidence for voxel-wise stability is present. Using the Human Connectome Project Data, we demonstrate how in a group analysis our method results in a higher number of selected clusters than when using only the conservative treshold. Further, we also find a larger overlap in the results than when only the liberal threshold is used.

5.1 Introduction

Following every scientific experiment the researcher is entrusted with weighting the cumulated evidence and inferring correct and relevant information. For Functional Magnetic Resonance Imaging (fMRI) data, signals are measured via the Blood Oxygen Dependant (BOLD). Evidence for brain activation is typically summarized in a statistical parametrical map (SPM) or a test image based on the general linear model (GLM) (e.g. Lindquist, 2008). These images consist of a summary statistic for each of the > 100.000 voxels, i.e. the small volumetric units that form the brain volume, and can be a summary of a single subject study, a group study (multi-subject study) or a meta-analysis. In each voxel, the evidence against the null hypothesis of no activation (H_0) is tested. When

H_0 is rejected, this provides evidence for the alternative hypothesis of true activation (H_1).

As the amount of voxels is huge, the decision process is a challenging endeavor. While not correcting for this multiplicity of tests would result in an excessive number of false positives (FP, see also Table 5.1), voxel-based corrections can result in overtly conservative results, especially when Bonferroni procedures are used (Worsley, 2007; Nichols & Hayasaka, 2003). A possible solution provided by cluster-based inference is the reduction of the dimensionality of the test problem (e.g. Forman et al., 1995). Woo, Krishnan, & Wager (2014) recently showed that the cluster-based corrections continue to be a popular choice for studies with fMRI data. The interpretation and validity of this correction however remain a recurring concern in the literature (Eklund, Nichols, & Knutsson, 2015; Woo et al., 2014; Hayasaka & Nichols, 2003).

		Decision	
		Conclude $\mathsf{H_0}$	Conclude $\mathsf{H_1}$
Truth	Active	False Negative (FN) Type II error	*True Positive* (TP)
	Inactive	*True Negative* (TN)	False Positive (FP) Type I error

Table 5.1 Table of events for Null Hypothesis Significance Testing (NHST) in which evidence against a null hypothesis H_0 is evaluated in the direction of an alternative hypothesis H_1.

In cluster-based inference, the feature of interest is a cluster, which is defined as a collection of neighbouring voxels that survive a first treshold T_{u1} (cluster-forming threshold) on a test image (Brett, Penny, & Kiebel, 2007). To determine the probability to observe a given cluster with size S under the H_0 of no activation, given this first threshold, one typically relies on a fast Random Field Theory based approximation (RFT; Friston, Holmes, Poline, Price, & Frith, 1996), although permutation-based alternatives exist (e.g. Holmes, Blair, Watson, & Ford, 1996; Nichols & Hayasaka, 2003). This RFT approach relies on several assumptions that are in practice very hard to verify, especially with regard to the smoothness, and the height of T_{u1} (Hayasaka & Nichols, 2003; Eklund et al., 2015).

In cluster-based testing, over 100 clusters are typically evaluated si-

multaneously. Two corrections for multiple testing are dominant in literature, Family-Wise error rate (FWE) control (e.g. Brett et al., 2007) and False-Discovery Rate (FDR) control (e.g. J. R. Chumbley & Friston, 2009; J. Chumbley, Worsley, Flandin, & Friston, 2010). FWE control uses the null distribution of the maximum cluster size $max(S)$ to control the probability of *at least* one false positive while FDR control uses the observed probability to control the number of false positive clusters among all discoveries (TP and FP).

Recently, Button et al. (2013) decribed how neuroscientific studies are often dealing with low power implying that important activation can be missed. Durnez, Roels, & Moerkerke (2014) also demonstrated this lack in power in studies using cluster-based inference. While allowing for more FP will increase power as the number of FN becomes smaller, this interplay between FP and FN in an fMRI data analysis is complicated by multiple testing corrections (Mumford, 2012; Durnez, Roels, & Moerkerke, 2014) and small sample sizes (Carp, 2012; Button et al., 2013). Lieberman & Cunningham (2009) have therefore argued for a better balance between FP and FN rates. They proposed a two-step procedure in which first a more lenient threshold is set on voxel-level (uncorrected α of 0.005). In a second step, only voxels that survive the first threshold and are part of a cluster with a size of at least 10 voxels, are retained. This procedure resembles cluster-based inference procedure in the sense that first a cluster-forming threshold is imposed and in a second step, decisions are based on cluster sizes. However, the choice of the thresholds is arbitrary and in contrast to cluster-based inference, is not based on a theoretical framework that enables formal FP control. The procedure could be criticized as an ad-hoc choice of parameters that enables to finetune the parameter configurations to obtain the desired results. Nevertheless, a voxel-wise evaluation of this approach showed that results are quite stable when compared to other voxel-wise testing procedures (Roels, Loeys, & Moerkerke, 2016).

Data-analytical stability can be measured through the variability on the number of selected features (e.g. voxels, clusters) when the same threshold for inference is used in an replication context (Qiu, Xiao, Gordon, & Yakovlev, 2006). Results that show a higher variation, can be considered as less stable. The concept was initially introduced in genetic association studies (e.g. Qiu et al., 2006; Gordon, Glazko, Qiu, & Yakovlev, 2007) but recently extended to fMRI (Durnez, Moerkerke, & Nichols, 2014; Roels, Moerkerke, & Loeys, 2015; Roels et al., 2016). Roels et al. (2016)

demonstrated that for group studies, voxel-wise FWE and FDR corrected analyses resulted in the same ROC curve and hence on average an equal trade-off between FP and FN (see also e.g. Durnez, Roels, & Moerkerke, 2014). However, FDR corrected analyses resulted in in a higher variability on the number of selected voxels (see also Durnez, Moerkerke, & Nichols, 2014; Qiu et al., 2006). We believe that data-analytical stability is an informative addition for the evaluation of test procedures, in which the current emphasis is mostly on average performance only.

The translation of the concept of stability to the fMRI context also resulted in the construction of voxel-wise re-selection rates (Roels et al., 2016). These rates allow to quantify the stability of a voxel in terms of reproducibility, given a fixed thresholding method, and have previously been added to the inference procedure within genetic association studies (Gordon, Chen, Glazko, & Yakovlev, 2009). Furthermore, as these rates can be computed for every thresholding method, they have the potential to add useful voxel-wise information to cluster-based inference. Indeed, one major restriction of cluster-based inference is the lack of a voxel-based interpretation of the results (Nichols, 2012; Woo et al., 2014; Durnez, Roels, & Moerkerke, 2014). As a significant cluster can only be interpreted in such a way that *in at least one voxel, somewhere in the cluster, there is non-zero signal* (Nichols, 2012; Poldrack, Mumford, & Nichols, 2011), the procedure becomes less attractive. By adding information on the data-analytical stability of a cluster, this can be partially resolved. As results of cluster-based inference have been shown to be unstable with low degrees of smoothness (Hayasaka & Nichols, 2003; Nichols, 2012), information of the stability may be advantageous in these situations. Quantification of this instability and adding this into the decision process may further improve cluster-based inference.

In this study, we propose a procedure along the lines of Lieberman & Cunningham (2009), given the good results with respect to stability found in Roels et al. (2016). First, we set a cluster-forming threshold that is typically used in cluster-based inference (uncorrected $\alpha = 0.001$). Next, we opt for a balance between two principled thresholds for cluster size. As opposed to Lieberman & Cunningham (2009), the cut-offs for cluster sizes are obtained through stochastic properties of clusters.

As FWE correction was found to be less variable than FDR corrections (Qiu et al., 2006; Durnez, Roels, & Moerkerke, 2014; Roels et al., 2016), we determine two FWE controlling thresholds for cluster size at cluster level: 1) a first relatively conservative threshold and 2) a more liberal

threshold. Finally, we add a measure of data-analytical stability into the decision process. Clusters of which the size falls between the two cluster size thresholds but demonstrate evidence of high stability, are retained.

In section *Method* we describe the proposed method in depth, in *Evaluation and Illustration of the method* we show how our procedure is evaluated using data from the Human Connectome Project (Van Essen et al., 2012). Next, we present the results and we conclude with a discussion.

5.2 Method

In this section, we first briefly describe the mass-univariate GLM approach to analyse fMRI data. Building on this voxel-wise GLM approach, Lieberman & Cunningham (2009) propose a two-step procedure in which they also incorporate a minimum required cluster size. Next, we propose a new method that incorporates cluster size in a more formal matter by deriving minimum cluster sizes based on RFT inference. Furthermore, voxel-wise data analytical stability is also taken into account.

5.2.1 Mass-univariate GLM

In a first stage, a GLM is fitted per subject (no index for the ease of notation) for the BOLD signal of each voxel over time $\mathbf{y_v}$ ($\mathbf{y_v} : y_{v1}, \dots y_{vn}, \dots y_{vN}$) with N: total number of time points, and with $v = 1, \dots, V$ the total number of voxels in the brain volume (see e.g. Kiebel & Holmes, 2007; Worsley et al., 2002; Poline & Brett, 2012).

$$\mathbf{y_v} = \mathbf{X}\boldsymbol{\beta_v} + \boldsymbol{\epsilon_v}, \tag{5.1}$$

In Equation (5.1) $\mathbf{X}$ is a matrix that represents the expected BOLD signal under brain activation. This is a convolution of the stimulus onset function(s) with the hemodynamic response function (HRF) (Henson & Friston, 2007) for the BOLD signal. $\boldsymbol{\epsilon_v}$ is the vector representing the residuals per voxel v. The estimands of interest are usually single parameters of the $\boldsymbol{\beta_v}$ vector or a linear contrast of several parameters within $\boldsymbol{\beta_v}$. These quantities are typically estimated based on weighted least squares estimation procedures that account for the temporal dependency (Cochrane & Orcutt, 1949; Kiebel & Holmes, 2007; Worsley et al., 2002).

In a single subject analysis, a T map or SPM is obtained based on these estimators and standard errors for each voxel. For group or multi-subject analyses, these estimators are transferred to the group level. Consider for

each voxel an estimator $\hat{\boldsymbol{b}}_{\boldsymbol{m}}$ ($m : 1 \ldots M$ with M the number of subjects) as the input contrasts for the group level (Beckmann, Jenkinson, & Smith, 2003). A GLM is used to weight the evidence over the M subjects (e.g. Mumford & Nichols, 2006):

$$\hat{\boldsymbol{b}} = \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{5.2}$$

where $\boldsymbol{X}_{\boldsymbol{M}}$ denotes the design matrix. After the estimation of γ (Beckmann et al., 2003; Mumford & Nichols, 2009; Worsley et al., 2002), a test statistic T is computed for each voxel.

To correct for multiple testing Lieberman & Cunningham (2009) propose a procedure that entails 2 thresholds to obtain a better balance between the number of FP and FN. For the first threshold these authors propose an uncorrected threshold T_u at the voxel level, i.e. $\alpha = 0.005$. Because under the null hypothesis of no activation this would result in an excessive amount of spuriously activated voxels, a second threshold at the cluster level is set. Only voxels that survive the first threshold and that lie within a cluster of a size of at least k voxels are retained. Lieberman and Cunningham (2009) propose to choose $k = 10$.

This procedure suffers from the fact that both thresholds are arbitrary determined and as such do not provide control on the FP rate in a principled way (Bennett, Wolford, & Miller, 2009). For example, the size of the volume is not taken into account by setting the thresholds nor is the minimum cluster size based on cluster-based inference.

5.2.2 Cluster-based Inference Including Data Analytical Stability at Voxel Level

Our proposed method to select clusters sets two thresholds and incorporates the average re-selection rate of the cluster. More specifically, a conservative and a liberal threshold are set based on the FWE corrected p values for cluster-based inference. Clusters that survive the first conservative threshold are selected. We stipulate however a *deliberation* for clusters that only survive the liberal threshold. For these clusters the re-selection rate needs to be sufficiently high to add to the selection.

RFT based thresholds

We define the thresholds using Random Field Theory (RFT) approximations for FWE corrected p-values (Worsley, 2007). RFT conveniently

allows to approximate the distribution of the extend of a cluster S as well as the distribution of the maxima: $max(S)$.

A cluster is defined as a collection of neighbouring voxels that exceeds a first threshold T_{u1}. The cluster extend in a Gaussian random field with dimension D can be re-formulated as $S \approx cH^{D/2}$ (Worsley, 2007) where H denotes the quadratic of height above treshold T_{u1} and where c equals:

$$\frac{FWHM^D T_{u1}^{D/2} P(T \leq T_{u1}|H_0)}{EC_D(T_{u1})\Gamma(D/2+1)} \tag{5.3}$$

with $EC_d(t)$ the d-dimensional Euler Characteristic density of the test statistic t and Γ the gamma function. The Full-Width Half Maximum ($FWHM$) describes the width of the smoothing kernel that should be applied on a dataset to achieve the same amount of smoothing in data.

The probability to obtain a cluster of size S under the H_0 of no activation for a given first threshold T_{u1} and a given $FWHM$ is approximated by

$$P\left(S > s|H_0\right) \approx exp\left(-T_{u1}\left(s/c\right)^{2/D}\right) \tag{5.4}$$

Based on these approximations, it is possible to derive FWE-corrected p values.

$$P\left(max(S) > s\right) \approx E(K)P\left(S > s\right) \tag{5.5}$$

with

$$E(K) \approx P\left(max(Z) > z\right) \approx \sum_{d=0}^{D} Resels_d \times EC_d(T_{u1}) \tag{5.6}$$

with $Resels_d$ the number of d-dimensional or resolution elements (Brett et al., 2007, p. 232) which can be estimated from the data. To select clusters based on the p-values derived in Equation (5.5), one defines a threshold α_{FWE}.

The formulation above clearly shows that the choice of the smoothing kernel has a critical impact on cluster inference. Spatial smoothing is considered as an essential pre-processing step to better comply with the RFT-assumptions (e.g. Hayasaka & Nichols, 2003). In Figure 5.1 the FWE corrected p-values are presented for a group analysis of 10 subjects over a hypothetical range of cluster sizes. In this analysis the data are smoothed with a kernel of 4 mm or 6 mm width. The derived smoothness (effective smoothness) then served as the input for the formulas above.

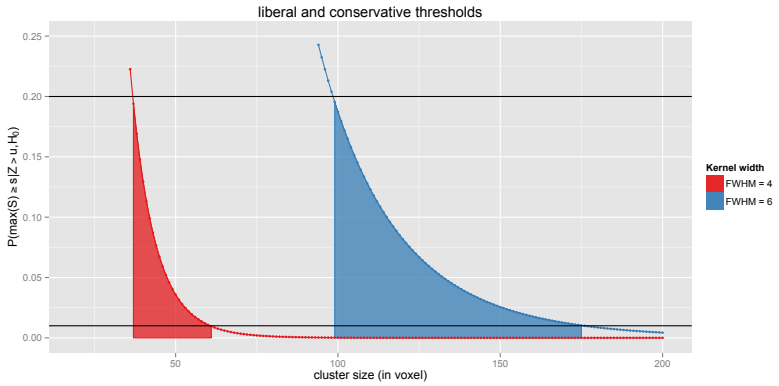


Figure 5.1 FWE corrected p-values in function of cluster size where the effective smoothness is based on an analysis of 10 subjects for an applied isotropic Gaussian smoothness kernel of 4mm and 6mm FWHM. The cluster forming treshold satisfies $P(T \geq T_{u1}) = 0.001$. The filled area is the *deliberation* zone between the *liberal* $\alpha_{FWE} = 0.2$ and *conservative* $\alpha_{FWE} = 0.01$ threshold.

Re-selection rate

For any given thresholding method, we define the re-selection rate (rr) at the voxel level in a bootstrap resampling context. From a dataset at hand we draw K bootstrap samples. For each of the bootstrap samples, we rerun the original analysis (thresholding method) and based on the image we determine the clusters that will be selected. Next, we create a binary map in which the voxels belonging to a selected cluster are set to 1. Voxels that do not belong to such clusters are set to 0. In the final step we sum op the K binary maps (with a value $k : 1, \ldots, K$ per voxel) and divide these through K, resulting in a proportion for each voxel that indicates the re-selection rate in that voxel over bootstrap samples. The average re-selection rate per cluster is obtained by averaging over all voxels within the cluster.

Procedure

We set the cluster forming threshold at $P(T \geq T_{u1}) = 0.001$ and compute the corresponding FWE corrected cluster p-values, accounting for smoothness and volume size. In a next step, we propose to set a *conservative* threshold at $\alpha_{FWE} = 0.01$, and the *liberal* threshold at a corrected

p value of $\alpha_{FWE} = 0.2$. Clusters that survive the conservative threshold are selected. Clusters that do not survive the liberal threshold are not selected. If a cluster only survives the liberal threshold, we let the cluster averaged re-selection rate be the determining factor. For the clusters that lie in the deliberation zone between the two thresholds (see also Figure 5.1) are only selected if the average is larger than $k/K = 2/3$, i.e. on average the voxels are selected in 2/3 of the bootstrap samples. The heuristic is outlined in Figure 5.2.

1. cluster survives the conservative threshold: **select the cluster**
2. cluster does not survive the liberal threshold: **do not select the cluster**
3. cluster only survives the liberal threshold: **select the cluster if average re-selection rate exceeds a pre-specified cut-off**

Figure 5.2 Scheme for the balanced multiple testing procedure.

5.3 Evaluation and Illustration of the method.

5.3.1 Evaluation

Real Data details

For the evaluation of our method, we use the 80 independent subjects package of the HCP (Van Essen et al., 2012) in a group analysis with samples of $n = 10$ or $n = 20$. We focus on 1 specific contrast of the emotion task, i.e. contrast 3 which compares faces versus shapes. Except for the applied Gaussian Kernel we use the minimal pre-processing protocol as described in (Glasser et al., 2013). For the smoothing we set a kernel width of 4mm and 6mm. We set the cluster-defining threshold T_{u1} so that it corresponds with an uncorrected p-value of 0.001 (Hayasaka & Nichols, 2003; Eklund et al., 2015; Woo et al., 2014). The FWE-corrected p values and the formation of the clusters are based on the FSL implemented command line tools (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012).

Setup

We compare 4 thresholding methods: 1) the conservative 0.01 FWE-corrected threshold ($\alpha_{FWE} = 0.01$); 2) the liberal 0.2 FWE-corrected threshold ($\alpha_{FWE} = 0.2$); 3) our above outlined balanced procedure with a liberal and conservative treshold, resp. $\alpha_{FWE} = 0.2$ and $\alpha_{FWE} = 0.01$ ($\alpha_{FWE}\&rr > 2/3$); and 4) a procedure for which only clusters are selected that survive the liberal $\alpha_{FWE} = 0.2$ and have an average re-selection rate that exceeds 2/3 ($rr > 2/3$). We add the fourth scenario to demonstrate that stability can also have a more prominent role in an inference strategy while ensuring that the FWE remains below a pre-specified level.

These four thresholding methods are juxtaposed on 4 criteria: 1) the number of selected voxels; 2) the number of selected clusters; 3) the overlap between the mutually independent samples; and 4) the number of overlapping voxels between the mutually independent samples. The number of selected voxels/clusters is based on the respective thresholded test image from all samples.

The overlap between mutually independent samples is determined using the Jaccard Index (Jaccard, 1901):

$$\omega = \frac{V_{j,l}}{V_j + V_l - V_{j,l}} \tag{5.7}$$

With $0 \leq \omega_{j,l} \leq 1$, and $V_{j,l}$ the number of voxels in the union both images j and l. V_j and V_l denote respectively the total amount of active voxels in image j & l. ω thus denotes the ratio of the total amount of voxels that are selected from both test images versus the amount of selected voxels in each of the images.

5.3.2 Illustration

The re-selection rate can be easily represented graphically by the method proposed by Allen, Erhardt, & Calhoun (2012). By adding the re-selection as the level of transparency to a SPM heat map, a more voxel-based interpretation is added to the difficult interpretation of cluster-based inference plots. Note that this illustration is only based on one sample from the HPC data with $n = 10$.

5.4 Results

Sample with $n = 10$

The results for the pairwise comparisons of the mutually exclusive samples with $n = 10$ can be found in Tables 5.2-5.3 for respectively a smoothing kernel of 4 and 6mm. In general, we find that the liberal FWE threshold of $\alpha = 0.2$ results in the highest number of selected clusters and voxels, while the conservative threshold of $\alpha = 0.01$ results in the smallest number of selected clusters and voxels. Obviously, we find that for the strategy that uses two thresholds, the number of selected clusters (and thus voxels) lies between these numbers of the two thresholds. Compared to FWE control on cluster level with $\alpha = 0.01$, we select more clusters but not at the cost of a substantial decline in the average overlap ω. On average we find about one half cluster extra if the data were smoothed with a 4mm width kernel while the difference is much smaller when a smoothing kernel of 6 mm was used.

Sample with $n = 20$

The results for the pairwise comparisons of the mutually exclusive samples with $n = 20$ can be found in Tables 5.4-5.5 for a smoothing kernel of 4 and 6mm respectively. In general we find a similar pattern as with a sample size of $n = 10$. We note however that with a smoothing kernel of 6 mm the average increase in selected voxels is about 100 when using the balanced criterion instead of an FWE control with $\alpha = 0.001$. In contrast to the sample size of $n = 10$, our procedure results on average in about 1.5 extra clusters, with a total additional number of voxels ranging up to on average about 100 voxels extra if a smoothing kernel width of 6 mm is used.

From Tables 5.2-5.5 it is interesting to note that the results based on the rr alone in combination with the liberal threshold a smaller amount of clusters is selected. However, a good overlap between the samples is found.

Illustration

The method and implementation of Allen et al. (2012) enables high dimensional visualisation of data properties. We use this principle to incorporate data analytical stability into classical fMRI plots using heat maps on brain slices. On Figure 5.3 we distinguish 3 different layers: 1) the classical heat map on the color scale with a Z-statistic; 2) the data analytical stability

Table 5.2 The average, median and standard deviation (sd) for ω, the number over overlapping voxels, the number of thresholded clusters and the number of thresholded voxels based on 100 mutually exclusive samples ($n = 10$) of the emotion task from the HPC data for an applied smoothing kernel of 4 mm.
$\alpha_{FWE} = 0.01$: thresholded clusters survive $\alpha_{FWE} = 0.01$; $\alpha_{FWE} = 0.2$: thresholded clusters survive $\alpha_{FWE} = 0.2$; α_{FWE} & $rr > 2/3$: thresholded clusters survive either 1) $\alpha_{FWE} = 0.01$ or 2)$\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained; $rr > 2/3$: thresholded clusters survive $\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained.

	ω			overlapping voxels			# clusters		# voxels	
Threshold	$\bar{\omega}$	median	sd	$\bar{n}$	median	sd	$\bar{c}$	sd	$\bar{v}$	sd
$\alpha_{FWE} = 0.01$	0.191	0.193	0.052	408.270	386.5	148.073	4.650	1.671	1267.610	467.057
$\alpha_{FWE} = 0.2$	0.188	0.185	0.050	420.110	395.5	147.905	6.495	2.729	1325.905	484.186
α_{FWE} & $rr > 2/3$	0.190	0.188	0.052	414.240	393.0	148.432	5.140	1.791	1283.935	464.606
$rr > 2/3$	0.211	0.209	0.055	451.111	423.0	181.526	4.638	1.504	1249.467	452.123

Table 5.3 The average, median and standard deviation (sd) for ω, the number over overlapping voxels, the number of thresholded clusters and the number of thresholded voxels based on 100 mutually exclusive samples ($n = 10$) of the emotion task from the HPC data for an applied smoothing kernel of 6 mm.
$\alpha_{FWE} = 0.01$: thresholded clusters survive $\alpha_{FWE} = 0.01$; $\alpha_{FWE} = 0.2$: thresholded clusters survive $\alpha_{FWE} = 0.2$; α_{FWE} & $rr > 2/3$: thresholded clusters survive either 1) $\alpha_{FWE} = 0.01$ or 2)$\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained; $rr > 2/3$: thresholded clusters survive $\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained.

	ω			overlapping voxels			# clusters		# voxels	
Threshold	$\bar{\omega}$	median	sd	$\bar{n}$	median	sd	$\bar{c}$	sd	$\bar{v}$	sd
$\alpha_{FWE} = 0.01$	0.269	0.273	0.063	1025.460	1010.0	325.259	3.460	1.480	2437.190	948.689
$\alpha_{FWE} = 0.2$	0.265	0.266	0.059	1069.280	1048.0	315.979	5.090	2.238	2570.845	976.736
α_{FWE} & $rr > 2/3$	0.269	0.270	0.062	1036.760	1017.0	320.570	3.695	1.589	2459.160	946.323
$rr > 2/3$	0.269	0.273	0.062	1014.430	929.5	322.304	3.230	1.395	2380.950	897.842

Table 5.4 The average, median and standard deviation (sd) for ω, the number over overlapping voxels, the number of thresholded clusters and the number of thresholded voxels based on 100 mutually exclusive samples ($n = 20$) of the emotion task from the HPC data for an applied smoothing kernel of 4 mm.
$\alpha_{FWE} = 0.01$: thresholded clusters survive $\alpha_{FWE} = 0.01$; $\alpha_{FWE} = 0.2$: thresholded clusters survive $\alpha_{FWE} = 0.2$; α_{FWE} & $rr > 2/3$: thresholded clusters survive either 1) $\alpha_{FWE} = 0.01$ or 2)$\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained; $rr > 2/3$: thresholded clusters survive $\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained.

	ω			overlapping voxels			# clusters		# voxels	
Threshold	$\bar{\omega}$	median	sd	$\bar{n}$	median	sd	$\bar{c}$	sd	$\bar{v}$	sd
$\alpha_{FWE} = 0.01$	0.426	0.423	0.045	2944.890	2970.0	355.236	6.640	2.520	4947.130	1016.831
$\alpha_{FWE} = 0.2$	0.418	0.418	0.043	2984.100	3006.0	358.814	10.440	3.618	5080.435	1033.630
α_{FWE} & $rr > 2/3$	0.420	0.420	0.043	2970.330	2989.5	357.117	8.315	3.110	5006.725	1021.758
$rr > 2/3$	0.421	0.421	0.045	2954.380	2973.5	360.862	6.130	3.190	4745.135	1003.038

Table 5.5 The average, median and standard deviation (sd) for ω, the number over overlapping voxels, the number of thresholded clusters and the number of thresholded voxels based on 100 mutually exclusive samples ($n = 20$) of the emotion task from the HPC data for an applied smoothing kernel of 6 mm.
$\alpha_{FWE} = 0.01$: thresholded clusters survive $\alpha_{FWE} = 0.01$; $\alpha_{FWE} = 0.2$: thresholded clusters survive $\alpha_{FWE} = 0.2$; α_{FWE} & $rr > 2/3$: thresholded clusters survive either 1) $\alpha_{FWE} = 0.01$ or 2)$\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained; $rr > 2/3$: thresholded clusters survive $\alpha_{FWE} = 0.2$ if the average $rr > 2/3$ is attained.

	ω			overlapping voxels			# clusters		# voxels	
Threshold	$\bar{\omega}$	median	sd	$\bar{n}$	median	sd	$\bar{c}$	sd	$\bar{v}$	sd
$\alpha_{FWE} = 0.01$	0.462	0.462	0.055	4834.550	4867.5	584.900	4.870	1.887	7704.285	1826.903
$\alpha_{FWE} = 0.2$	0.452	0.453	0.051	4947.440	4971.5	589.669	8.100	2.637	8001.135	1864.993
α_{FWE} & $rr > 2/3$	0.456	0.457	0.053	4897.310	4952.0	593.906	6.115	2.476	7820.275	1849.931
$rr > 2/3$	0.457	0.458	0.054	4825.790	4829.5	597.160	4.605	2.622	7365.595	1746.474

displayed as the transparency of the colors and 3) the contours indicate the selected clusters. We demonstrate the principle for the analysis for one sample with $n = 10$, using a smoothing kernel of respectively 4 mm and 6 mm width.

As two analyses only differ in the amount of smoothing applied, the results in Tables 5.6 and 5.7 highly resemble each other, which is also noticeable in the left and the middle panel from Figure 5.3. Based on the results in Table 5.6, next to selection of clusters 5.6.1-5.6.3, we additionally select cluster 5.6.5, which has a $p = 0.079$ but an average re-selection rate of 0.827. It is furthermore remarkable that with a smoothing kernel of 6 mm, a similar pattern occurs, in which cluster 5.7.4 is added to the selection. Note however the substantial difference in size due to the choice of kernel width.

The visualisation in Figure 5.3 also allows to explore the data analytical stability in specific regions. For the selected clusters, we see that those at the borders are less stable, especially with a kernel width of 4 mm. Furthermore, Figure 5.3 allows to inspect the data analytical stability of voxels that do not exceed the that do not exceed the cluster-forming threshold. Based on these figures, we can seem both for 4 and 6 mm smoothing, relatively stable activation in the left temporal cortex.

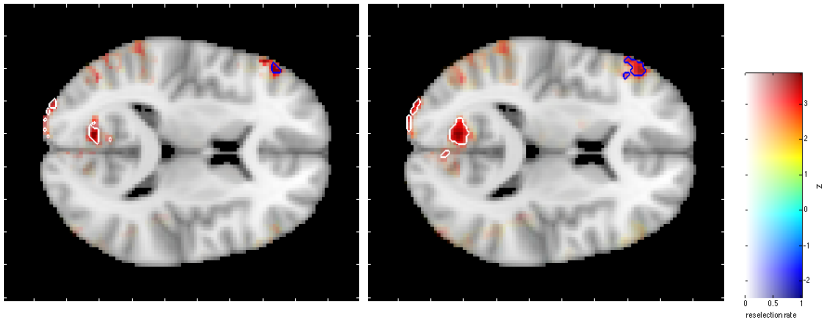


Figure 5.3 Activation for contrast 3 in the HCP data. The white contours indicate clusters that are selected based on cluster-based inference with $\alpha_{FWE} = 0.01$. The blue contour indicates a cluster that only survives the liberal threshold, but also has an average $rr > 2/3$. The colors correspond with the value of the Z statistic, more transparency indicates a lower re-selection rate.

Table 5.6 Results from 1 FSL-based cluster analysis for $n = 10$ with a smoothing kernel of 4 mm.

	ID	size	$maxZ$	FWE p	X	Y	Z	average re-selection rate
	5.6.1	790	4.750	< 0.0001	31	17	33	0.724
survive $\alpha_{FWE} = 0.01$	5.6.2	375	5.330	< 0.0001	66	21	30	0.774
	5.6.3	64	4.800	0.001	64	36	27	0.836
	5.6.4	31	4.130	0.058	36	63	28	0.510
survive $\alpha_{FWE} = 0.2$	**5.6.5**	29	4.010	**0.079**	20	83	41	**0.827**
	5.6.6	26	3.740	0.127	56	29	39	0.365

Table 5.7 Results from 1 FSL-based cluster analysis for $n = 10$ with a smoothing kernel of 6 mm.

	ID	size	$maxZ$	FWE p	X	Y	Z	average re-selection rate
	5.7.1	1769	4.740	< 0.0001	31	17	34	0.745
survive $\alpha_{FWE} = 0.01$	5.7.2	850	5.140	< 0.0001	64	36	27	0.805
	5.7.3	195	4.330	< 0.0001	34	63	28	0.624
survive $\alpha_{FWE} = 0.2$	**5.7.4**	112	4.060	**0.013**	20	83	41	**0.700**

5.5 Discussion

We presented a new method for the selection of clusters that balances between a conservative and a liberal threshold for multiple comparison correction and contains data analytical stability. We retain clusters that either survive the conservative threshold or either survive the liberal threshold but have a high average re-selection rate. Our results show that this procedure is succesful in the selection of more clusters and, consequently more voxels are declared active without substantial losses performance compared to similar thresholding method.

While the proposal of Lieberman & Cunningham (2009) also aimed at a better balance between FP's and FN's, it lacks theoretical motivation for such balance. By setting a theoretical framework to define the two thresholds our proposal intrinsically takes into account the properties of the data (i.e. the size of the volume, the smoothness and the height of the first thresholds). Secondly, the addition of a voxel-wise metric within the cluster-based inference increases the interpretability of the voxels *in* a cluster. Indeed, in the illustration of the results it provides useful information in two ways. First, within a selected cluster it allows to differentiate between within-cluster regions that are highly stable versus regions that not stable. Second, based on the re-selection rates, we can detect regions that even do not survive liberal threshold but are relatively stable.

While the concept of data analytical stability has only been recently introduced in the analysis of fMRI data (Roels et al., 2015, 2016) for the evaluation of methodological choices, it has previously been used to improve corrections for multiple testing (e.g. Gordon et al., 2009) also in fMRI (Durnez, Roels, & Moerkerke, 2014). The implementation of data analytical stability in the voxel-wise single-subject analysis of fMRI was succesful in the improvements of stability of the FDR correction for multiple testing (Durnez, Roels, & Moerkerke, 2014). It was however conceived as very computationally intensive, especially since the balance between FP and FN was not improved on average. Although our method is also computationally intensive, it does assure that only the moste stable clusters are selected from the twilight zone between the thresholds. Furthermore, our methods is based on the FWE correction for multiple testing previously described as more stable than the FDR correction (Qiu et al., 2006; Durnez, Moerkerke, & Nichols, 2014; Roels et al., 2016).

As the inferential validity of cluster-based inference might not be guaranteed for low first thresholds or too low smoothing (Eklund et al., 2015;

Hayasaka & Nichols, 2003), we insure this by setting a high first threshold ($\alpha = 0.001$) and incorporating the data analytical stability. Our method is however sufficiently flexible to use permutation-based tresholds, in contrast to the currently used RFT-based thresholds. Moreover, the inclusion of the average re-selection of the voxels within a cluster provides a useful metric for the potential instability of the cluster-based tresholds in that with low intrinsic smoothing.

At last we stress that the proposed procedure is relatively immune to so-called *p*-hacking, the coarse practice to smuggle in significance of results that lie near a fixed threshold. With the choice and advice for a resolute cluster-forming threshold and the a priori choice of two thresholds, we stress the importance of avoiding post-hoc manipulation.

Conclusion In this paper we have presented a procedure that successfully balances between sufficient control on the amount of false positives and sufficiently high power by including data analytical stability in cluster-based inference. The inclusion of data analytical stability additionally aids the interpretation of the results.

References

Allen, E. a., Erhardt, E. B., & Calhoun, V. D. (2012). Data Visualization in the Neurosciences: Overcoming the Curse of Dimensionality. *Neuron*, *74*(4), 603–608.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, *20*(2), 1052–63.

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, *4*(4), 417–22.

Brett, M., Penny, W., & Kiebel, S. (2007). Parametric procedures. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Button, K. S., Ioannidis, J. P. a., Mokrysz, C., Nosek, B. a., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, *14*(5), 365–76.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Chumbley, J., Worsley, K., Flandin, G., & Friston, K. (2010). Topological FDR for neuroimaging. *NeuroImage*, *49*(4), 3057–64.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, *44*(1), 62–70.

Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61.

Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, *84*, 45–64.

Durnez, J., Roels, S., & Moerkerke, B. (2014). Multiple testing in fmri: a case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, *56*(4).

Eklund, A., Nichols, T., & Knutsson, H. (2015). Can parametric statistical methods be trusted for fMRI based group studies? *ArXiv e-prints*.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fmri): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647.

Friston, K. J., Holmes, a., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, *4*(3 Pt 1), 223–235.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., ... Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Gordon, A., Chen, L., Glazko, G., & Yakovlev, A. (2009). Balancing Type One and Two Errors in Multiple Testing for Differential Expression of Genes. *Computational statistics & data analysis*, *53*(5), 1622–1629.

Gordon, A., Glazko, G., Qiu, X., & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, *1*(1), 179–190.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Henson, R., & Friston, K. J. (2007). Convolution models for fmri. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 193-210). Elsevier Ltd./Academic Press.

Holmes, A. P., Blair, R. C., Watson, J. D., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, *16*(1), 7–22.

Jaccard, P. (1901). Distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547-579.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Kiebel, S., & Holmes, A. P. (2007). The general linear model. In K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (chap. 8). Elsevier Ltd./Academic Press.

Lieberman, M. D., & Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, *4*(4), 423–8.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464.

Mumford, J. A. (2012). A power calculation guide for fMRI studies. *Social cognitive and affective neuroscience*, *7*(6), 738–42.

Mumford, J. A., & Nichols, T. E. (2006). Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 42–51.

Mumford, J. A., & Nichols, T. E. (2009). Simple group fMRI modeling and inference. *NeuroImage*, *47*(4), 1469–75.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–5.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional mri data analysis*. New York: Cambridge University Press.

Poline, J.-B., & Brett, M. (2012). The general linear model and fMRI: Does love last forever? *NeuroImage*, *62*(2), 871–880.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, *7*.

Roels, S. P., Loeys, T., & Moerkerke, B. (2016). Evaluation of Second-Level Inference in fMRI Analysis. *Computational Intelligence and Neuroscience*, *2016*(Article ID 1068434), 22.

Roels, S. P., Moerkerke, B., & Loeys, T. (2015). Bootstrapping fmri data: Dealing with misspecification. *Neuroinformatics*, *13*(3), 337-352.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, *91*, 412–9.

Worsley, K. J. (2007). Random field theory. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 232-236). Elsevier Ltd./Academic Press.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*(1), 1–15.

6 General Discussion

6.1 Summary of the Present Work

In Chapter 2 we have presented a study on the properties of bootstrap procedures in fMRI. Although *non-parametric* strategies for inference have previously been applied in the fMRI literature, only few studies have investigated the performance of these techniques. In the first study we have investigated both the inferential properties and the ability to mimic properties of the original sample. In general we found that bootstrap procedures perform well under various conditions and can cope with some mis-modeling.

In the second study, presented in Chapter 3, we have introduced data analytical stability in the analysis of cluster-based inference. Data analytical stability was assessed while the distinction with validity and reliability was maintained. We demonstrated the capabilities of this assessment in single-subject studies in both real and synthetic data.

In the third study (Chapter 4) we implemented data stability in a voxel-based context and evaluated several choices in the analysis of multi-subject data. Again, using data analytical stability metrics, we were able to distinguish stable from less stable strategies. In this study we also introduced the concept of re-selection rates, which is defined in a replication context. These re-selection rates indicate the degree to which voxels are selected over bootstrap samples.

In the fourth study (Chapter 5), we have implemented data analytical stability *in* the inference strategy. For cluster-based inference, we let the per-cluster averaged re-selection rates be decisive in the selection of active clusters. In a procedure with two thresholds we demonstrated the usefulness of this metric. Furthermore, the inclusion of these re-selection rates allowed for a more voxel-wise interpretation of the detected clusters.

In this section, we will further elaborate on 3 important topics that were put forward in this dissertation: the inference strategy, the role of stability in the evaluation of methodological choices, and the role of stabil-

ity *in* the inference strategy. At last, we consider possible future research topics that relate to the studies of this dissertation.

6.2 Inference Strategy

Throughout the first three studies the choice of inference strategy was a recurrent theme, especially because of its importance in the decision process. While in the first study we explored the properties of bootstrap, in the second and the third study we contrasted parametric and permutation-based inference strategies.

Although it is well-known that the validity of parametric inference strategies is bound by the degree to which the underlying assumptions are guaranteed, in practice there is little verification in fMRI studies although tools are available (Luo & Nichols, 2003; Zhang, Luo, & Nichols, 2006). The improper use of parametric inference has furthermore been demonstrated to result in an uncontrolled modeling of e.g. the noise in single-subjects analyses (Lenoski, Baxter, Karam, Maisog, & Debbins, 2008) and in invalid inference (Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012). Even though in our first study we demonstrated the potential of bootstrap procedures, we found that accounting for the temporal structure is crucial. More specifically, the correspondence between an explicit parametric temporal model and the temporal pattern in the data is a critical condition for the bootstrap to perform well. If this temporal model does not correspond to the pattern in the data, it can be further improved with additional temporal grouping of observations (blocking). These findings are in line with earlier findings on performance of such temporal models for fMRI data in parametric inference (Worsley et al., 2002; Lenoski et al., 2008).

While the flexibility of bootstrap procedures, e.g. to use a GLM, is generally seen as an advantage (see e.g. Nichols & Hayasaka, 2003), it remains vulnerable on the specification of the mean signal in such models. Although we varied the shape of the BOLD signal globally, our study did not account e.g. for regional differences in the shape of the BOLD signal. Also, concerning the (global) shape of the hemodynamic responce, more flexible models such as finite impulse response (see e.g. Henson & Friston, 2007) can also be used within the bootstap framework. At last, from our simulations it should be clear that in the presence of heavy noise bootstrap procedures cannot result in valid inferential conclusions, but neither can parametric approaches.

In general, we also found good performances of permutation-based inference. Our simulation study in Chapter 4 e.g. did not reveal large differences beween permutation and parametric inference with respect to the balance between false positives and true positives in voxel-wise analyses. However, notwithstanding that FDR corrections for multiple testing were generally found to be more variable, this pattern was more explicit in permutation-based inference. One possible explanation is that, by default, the p-values are derived from only 5000 permutations, while in our example in total $2^{15} = \pm 32.000$ permutations are possible, making the p-values approximative.

While our implementation of data analytical stability in Chapter 5 only used parametrically derived thresholds, this could easily be extended to permutation inference. Indeed, via permutation-based inference it is also possible to arrive at a FWE corrected thresholds by using the maximum test statistic (see e.g. Nichols & Hayasaka, 2003).

It has been shown very recently that the fast computation of FWE thresholds based on RFT based parametric inference may result in an uncontrolled amount of false positives if e.g. the cluster-defining threshold is set too low (Eklund, Nichols, & Knutsson, 2015). These authors used resting-state data (no activation expected) to derive inferential properties. Using a wide range of fictive designs and smoothing kernels they found spurious activation with RFT based inference, but not with permutation-based inference. Together with recent advances in the development of permutation methods for neuroimaging data in a wide range of settings (Winkler, Ridgway, Webster, Smith, & Nichols, 2014; Winkler et al., 2016), this makes it an attractive alternative for parametric inference.

Despite the fact that assumptions are not systematically checked, pure parametric inference procedures continue to be play a very important role in fMRI data analysis. In general we found good performance for alternatives methods. It should however be noted that these alternatives are no cure-all for junk data or data that was improperly pre-processed. However, given acceptable levels of noise that cannot be modelled parametrically, bootstrap and permutation can offer welcome alternatives. Furthermore, with the recent advances in computational speed and parallell processing, we are hopeful for a more widespread use of permutation and bootstrap-based inference. In this perspective, our findings contribute to the relatively small literature on fMRI and bootstrap inference[1] (see e.g. Darki

[1] An explorative search on Web-of-science [2016/02/23] resulted in 38.101 research articles with the search term "fMRI", of wich only 100 were found when adding "boot-

& Oghabian, 2013).

6.3 Stability as an Evaluation Criterion

Throughout Chapter 3 and 4 we have assessed data analytical stability for respectively clusters and voxels. In both studies we quantified the variability on the selected features and used this as an evaluation criterion. While using both synthetic and real data, we distinguished stability from the validity and the reliability of the results. While we stress the importance of considering all three aspects in a methodological evaluation, stability added useful information. Via the real data application we furthermore demonstrated that stability can be assessed in every study, and can ultimately be applied as an additional selection criterion (see Section 6.4).

To illustrate the use of stability in an evaluation, we take back the case of adaptive smoothing in Chapter 3. It was contrasted with Gaussian smoothing for a range of kernel widths in parametric and permutation inference. When using permutation inference for cluster sizes we found a (relatively) better distribution of the p-values under H_0 of no activation compared to the distribution obtained via parametric inference. This is indicative for (relatively) more valid results with no large difference due to the type of smoothing. Next, when considering the reliability of the results, in general, we found that using adaptive smoothing was advantageous compared to classical Gaussian smoothing. However, when permutation-based inference was used in combination with such adaptive smoothing, the use of larger kernel widths resulted in more variability on the results. It is due to the addition of the perspective of data analytical stability that these differences emerge.

Similarly, when comparing FDR and FWE as methods for voxel-wise multiple comparisons correction in Chapter 4, we found interesting properties of stability. While the balance between false positives and true positives was found to identical for FDR and FWE, there were clear differences in stability of these methods. It is important to stress here, that this finding was, in general, irrespective of the amount of voxels that are selected. For the end user, confronted with the choice for a strategy to correct for multiple testing, this is an important consideration. As both methods are on the same balance between true and false positives, it might for example

strap" to the query, when combining "fMRI AND permutation" this resulted in 182 research articles.

be better to opt for a more leniently FWE corrected threshold. Although this method is typically conceived as more conservative, it can be made equally lenient as FDR while it results in more stable results (see also Durnez, Roels, & Moerkerke, 2014).

The quantification of stability, i.e. the variability on the amount of selected features, makes the use of clusters as feature of interest challenging. Due to the fact that clusters consist out of voxels that first have to exceed a threshold, (re-sampled) samples might consist out of a different number of clusters. This is in sharp contrast with voxels, where the total amount of voxels is always bounded by V. Although clusters can be described by their mass, peak or extend, it remains relatively difficult to characterize the stability of a cluster (Nichols, 2012), because of their intrinsic spatial properties (which are also the reason for the large popularity of this type of inference, see e.g. Woo, Krishnan, & Wager, 2014).

Next to the current efforts to improve the reproducibility via better reporting methods (after the eye-opening study of Carp, 2012), the investigation of the stability of the methodological choices per se is an important assessment. Via the proposed quantification it serves as a proxy of reproducibility. In this light, we have contributed to the investigation of the impact of methodological choices in fMRI data analysis. We are hopeful that the assessment of stability can further improve these choices, or at least alert researchers when no satisfying evidence for reproducible results can be attained.

At last, with the use of the re-selection rates, introduced in Chapter 4 and further implemented in 5, applied stability in a practical way. We further elaborate on this in the next section.

6.4 Stability in the Decision Process

In Chapter 5 we implemented data analytical stability in the decision process to improve it. With the voxel-wise re-selection rates, introduced in Chapter 4, we let the selection of clusters depend on the these rates.

As introduced before, due to the intrinsic spatial characteristics of clusters, it is difficult to assess the stability of a selected cluster in a replication context. This is why we opted to average the re-selection rates of the constituent voxels. We have implemented the stability in a procedure that sets 2 threshold to allow for balanced results. Clusters that are not selected based on a first threshold can be selected if the average re-selection rate is sufficiently large and if it exceeds a second threshold. While such

procedure neatly illustrates the added value of the quantification of stability, it could equally well have been added to procedures that only use one threshold.

In the same regard, although we introduced this strategy in the context of cluster-based inference, it can easily be transformed to settings with voxel-wise corrections for multiple testing. Woo et al. (2014) for example advise to use voxel-wise corrections for multiple testing (compared to cluster-based) if sufficient power can be ascertained. In these cases it could be interesting to set 2 fixed voxel-wise defined FWE thresholds, either based on parametric or permutation inference. We found furthermore that the results based on such FWE thresholds are stable and can result thus in more stable conclusions. With the addition of data analytical stability, we can also balance between less false negatives and slightly more false positive in a principled way.

We note that our work is complementary with earlier work in resting-state data analysis (see e.g. Bellec, Rosa-Neto, Lyttelton, Benali, & Evans, 2010). While in this field a larger focus lies on multi-variate approaches, we focused exclusively on mass-univariate approaches for modeling.

6.5 Future Research

Implementation One of the developments that are potentially very important for a better assessment of stability is the further development of a homogeneous ecosystem for the analysis of neuroimaging data. Although there is wide range of freely available software, such homogeneous ecosystem can further improve systematic evaluations. For example, the 3 most used software packages, i.e. SPM (Ashburner, 2012), FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) and AFNI (Cox, 1996), all have their own ecosystem. Of course, the developments in NiPy and NiPype (resp. Millman & Brett, 2007; Gorgolewski et al., 2011) cannot be underestimated and the value of the the translation of existing code into one single environment is great. On the other hand however, a substantial amount of statistical improvements is still implemented in the statistical language **R** (R Core Team, 2015). One single ecosystem that incorporates the lastest techniques in data manipulation, statistics and visualisation is something that can lead to a better evaluation of pipelines.

Combination of inference strategies Given its flexibility, bootstrap based confidence intervals could be added to the inference phase easily. In this way, these confidence intervals could serve as an alternative to describe the single subject variability. More specifically, if there is misspecification in the degree of variability on the first level, this will affect the construction of parametrically based confidence intervals. Although the use of this kind of intervals has been proposed quite some time ago (see e.g. Biswal, Taylor, & Ulmer, 2001), it has however not often been used since then. Also, the use of (bootstrap-based) confidence intervals can allow for a better understanding of the size of the *effect* that has been detected in either first- or second-level analyses.

The use of real data in the evaluation of methods Some recent evaluation studies have used (large) samples to study methodological choices (Eklund et al., 2012; Eklund et al., 2015). As simulations might suffer from being sufficiently realistic, this can result in biased conclusions and complicate further theorizing (Welvaert & Rosseel, 2012). This is why the availability of such large datasets is crucial for both the development and evaluation of new methods, but also for the continuous evaluation of existing methods. For example, in our latest study we demonstrated the abilities of the our proposed methodology on real data. This would not have been possible without the recent massive efforts on making data available (see e.g. Poline et al., 2012; Van Essen et al., 2012; Poldrack & Gorgolewski, 2014). The current attention for data sharing and making large datasets widely available is a necessary step forward towards more validation based on real data.

Golden standard or better methods reporting? One of the most interesting topics is whether the scientific community should work to a golden standard for the analysis of fMRI data to limit the methodological variability. On the other hand, is better methods reporting a sufficient recommendation to avoid this variability? With recent initiatives such as the "Committee on Best Practices in Data Analysis and Sharing" from the organisation of human brain mapping, a better reporting is promoted extensively. We also believe that this might lead to a more considerate selection of methodologies. While a golden standard methodology might result in incautious use of methodologies, this can be a good starting point for every analysis. In this regard, an additional motivation for deviating from the standard can further justify choices. With additional

motivation, methodological choice "A", rather than choice "B" which is e.g. more spatially accurate, can be rationalized. This could further encourage researchers to make even more scrutinized methodological choices and could eventually result in better scientific practice.

6.6 Conclusion

In this doctoral dissertation we have implemented and assessed data analytical stability in the analysis of brain activation data in fMRI studies. We used this quantification to assess the reproducibility within a replication context. While in the first study we investigated the necessary conditions for a good replication context, in the second and the third study we used stability to investigate the impact of methodological choices. In the fourth study we implemented data analytical stability as a selection criterion for activation. With this dissertation we hope to have raised more awareness on the quantification of stability of results in the complex analysis of fMRI data.

References

Ashburner, J. (2012). SPM: A history. *NeuroImage*, *62*(2), 791–800.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–1139.

Biswal, B. B., Taylor, P. a., & Ulmer, J. L. (2001). Use of jackknife resampling techniques to estimate the confidence intervals of fMRI parameters. *Journal of computer assisted tomography*, *25*(1), 113–20.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and biomedical research, an international journal*, *29*(3), 162–173.

Darki, F., & Oghabian, M. A. (2013). False positive control of activated voxels in single fMRI analysis using bootstrap resampling in comparison to spatial smoothing. *Magnetic resonance imaging*, *31*(8), 1331–1337.

Durnez, J., Roels, S., & Moerkerke, B. (2014). Multiple testing in fmri: a case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, *56*(4).

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Eklund, A., Nichols, T., & Knutsson, H. (2015). Can parametric statistical methods be trusted for fMRI based group studies? *ArXiv e-prints*.

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, *5*(13), 1-15.

Henson, R., & Friston, K. J. (2007). Convolution models for fmri. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (p. 193-210). Elsevier Ltd./Academic Press.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Lenoski, B., Baxter, L. C., Karam, L. J., Maisog, J., & Debbins, J. (2008). On the performance of autocorrelation estimation algorithms for fmri analysis. *IEEE Journal of Selected Topics in Signal Processing*, *2*, 828-838.

Luo, W.-L., & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, *19*, 1014-1032.

Millman, K. J., & Brett, M. (2007). Analysis of Functional Magnetic Resonance Imaging in Python. *IEEE Computing in Science & Engineering*, *9*(3), 52–55.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–5.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419–46.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, *17*(11), 1510–1517.

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., ... Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*, 9.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., ... Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231.

Welvaert, M., & Rosseel, Y. (2012). How ignoring physiological noise can bias the conclusions from fMRI simulation results. *Journal of neuroscience methods*, *211*(1), 125–32.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–97.

Winkler, A. M., Webster, M. A., Brooks, J. C., Tracey, I., Smith, S. M., & Nichols, T. E. (2016). Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping*, *37*(4), 1486-1511.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, *91*, 412–9.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*(1), 1–15.

Zhang, H., Luo, W.-L., & Nichols, T. E. (2006). Diagnosis of single-subject and group fMRI data with SPMd. *Human brain mapping*, *27*(5), 442–51.

7 Nederlandstalige Samenvatting

De afgelopen 25 jaar werd gekenmerkt door een enorme toename in het aantal studies dat gebruikmaakt van functionele Magnetische Resonantie Imaging (fMRI). Deze beeldvormingstechniek laat toe om activatie in de hersenen te meten aan het hand van het Blood Oxygen Level Dependent (BOLD) signaal. Dit signaal is gebaseerd op de magnetische eigenschappen van zuurstofarm en zuurstofrijk bloed. Op deze manier kan men *in vivo* beelden van de hersenen met een MRI scanner maken. De populariteit van de techniek is zowel te wijten aan haar niet-invasieve karakter als aan de mogelijkheid om de hersenactiviteit te registeren over de tijd met een grote spatiale accuraatheid.

Na het ontwikkelen van een goed studieopzet voor de onderzoeksvraag en de registratie van de data, wordt de data geanalyseerd. Bij deze analyse worden de over de tijd gescande beelden ingedeeld in kleine kubussen, voxels genaamd. Per voxel is er een BOLD-signaal gemeten over de tijd. Een volledig hersenvolume bestaat bovendien uit meer dan 100.000 voxels. We onderscheiden 4 opeenvolgende fases in de analyse: 1) het opzetten van de studie; 2) de *pre-processing* waarin de data wordt voorbereid voor de verdere analyse; 3) het modelleren van het signaal; en 4) de fase waarin conclusies worden geïnfereerd vanuit de data.

Bij de verzameling van de data worden typisch verschillende subjecten gerekruteerd. In dit proefschrift bestuderen we de analyse van experimenten waarbij subjecten een taak uitvoeren. Het doel van dergelijke studie is dan om te zien in welke hersenregio's er activatie is. Via een *mass-univariate* benadering wordt er per voxel een model opgesteld voor de analyse van de data (fase 3). Na het modelleren van de data binnen één subject, gebeurt een analyse over subjecten heen. Hoewel de veralgemening van de resultaten over subjecten meestal het doel van een studie is, zijn individuele analyses geen uitzondering. Uiteindelijk wordt op basis van een statistisch model de evidentie in een toetsstatistiek omgezet. Op deze manier kan men dan bepalen of er activatie heeft plaatsgevonden in een welbepaalde voxel of regio (met spatiale locatie). De modellen worden

simultaan geschat in de 100.000 voxels. Hierdoor vergroot de kans op fouten bij het infereren van conclusies uit de data. Er werden in de literatuur verschillende correcties uitgewerkt op basis van kenmerken van voxels of van groepen van voxels (clusters).

In de context van fMRI analyses toonde Carp (2012) de grote diversiteit aan binnen het gebruik van methoden (in fases 1,2,3 en 4). Deze grote variabiliteit heeft mogelijk ook een invloed op de reproduceerbaarheid van de gegevens. Een reproduceerbaarheid die niet hoog wordt geschat indien men hiervoor schattingen van de betrouwbaarheid gebruikt. Het blijft echter een moeilijke opdracht om de berekening van reproduceerbaarheid in kaart te brengen, hoewel het een hoeksteen van de wetenschappelijke praktijk is (Bennett & Miller, 2013).

In dit proefschrift schuiven daarom het concept van data-analytische stabiliteit naar voor. Het werd oorspronkelijk geïntroduceerd in context van de statistische analyses van genoomdata (Qiu, Xiao, Gordon, & Yakovlev, 2006). Het belang aan van data-analytische stabiliteit werd aangetoond door de finale selectie van genen als een random variabele in een replicatiecontext te beschouwen. Hierdoor kan de variabiliteit op deze selectie berekend worden. We beogen het concept van data-analytische stabiliteit te introduceren bij de analyse van fMRI data. Resultaten die een hoge stabiliteit hebben worden gekenmerkt door een lage variatie in het aantal geactiveerde voxels of clusters. Om de stabiliteit te bepalen maken we gebruik van een replicatiecontext, hierdoor wordt het een proxy van reproduceerbaarheid. Door gebruik te maken van *re-sampling* techniek, het herhaaldelijk trekken van steekproeven uit een geobserveerde dataset, en door gebruik te maken van gesimuleerde data kunnen we de data-analytische stabiliteit berekenen op basis van respectievelijk echte data en synthetische data.

In de eerste studie onderzoeken we hoe we met bootstrapping, een re-sampling techniek, op een goede manier nieuwe steekproeven kunnen trekken uit de geobserveerde data. Enerzijds laat bootstrapping toe om conclusies te trekken op basis van de data (fase 4), anderzijds vormen deze procedures een essentieel onderdeel voor de berekening van data-analytische stabiliteit. Aan de hand van verschillende gesimuleerde scenario's, met verschillende vormen van signaal en ruis, evalueren we hoe bootstrapprocedures presteren. Meer specifiek onderzochten we hierbij de invloed van het al dan niet correct modelleren van de ruis en/of het signaal. Op basis van deze studie besluiten we dat bootstrapprocedures die gebruik maken van minimale assumpties tot een goed behoud van de da-

takarakteristieken kunnen leiden. Indien er bovendien gegevens beschikbaar zouden zijn over de structuur van de ruis op de data, dan kan deze informatie gebruikt worden en leiden tot het trekken van goede conclusies. Op basis van deze studie kunnen we besluiten dat bootstrapping binnen een waaier van scenarios tot goede resultaten leidt voor zowel het trekken van conclusies als het behoud van kenmerken van de steekproef.

In de tweede studie conceptualiseren en implementeren we data-analytische stabiliteit binnen de statistische analyse van de data van één subject. Wij voegen in deze studie data-analytische stabiliteit toe als criterium in de evaluatie van een methode. We argumenteren in deze paper dat naast de validiteit van de methode (wordt een verwacht aantal statistische fouten gemaakt?) en naast de betrouwbaarheid (stemt het activatiepatroon overeen met de ware activatie?) er eveneens oog moet zijn voor de stabiliteit van de resultaten. Concreet onderzoeken we keuzes binnen de pre-processing (fase 2) en binnen de manier om conclusies te trekken op basis van clustereigenschappen (fase 4). Bij de vergelijking van deze keuzes vinden we inderdaad evidentie voor een toegevoegde waarde van data-analytische stabiliteit.

In de derde studie implementeren we het concept van data-analytische stabiliteit binnen de analyse van meerdere subjecten. We evalueren mogelijkheden bij het modelleren van de data van meerdere subjecten (fase 2) en bij het trekken van conclusies op basis van voxelgebaseerde inferentiemethodes (fase 4). Door in de evaluatie ook data-analytische stabiliteit op te nemen, vinden we verschillen tussen inferentiemethodes met dezelfde validiteit. Dit levert extra evidentie voor de waarde van data-analytische stabiliteit. In deze studie introduceren we ook de voxelgebaseerde herselectieratio's. Dit is een maat gebaseerd op data-analytische stabiliteit, meer bepaald is dit de mate waarin een voxel geselecteerd wordt in een replicatiecontext (bijv. aan de hand van bootstrap steekproeven).

In de vierde studie gaan we na hoe we data-analytische stabiliteit kunnen gebruiken in de analyse van fMRI data. Dit doen we door data-analytische stabiliteit, gekwantificeerd aan de hand van de herselectieratio's, mee te nemen in een procedure voor de selectie van significante clusters. Hiervoor baseren we ons enerzijds op de bevinding dat clustergebaseerde inferentie op zich niet altijd tot goede beslissingen leidt (bijv. Hayasaka & Nichols, 2003; Eklund, Nichols, & Knutsson, 2015) en anderzijds op onze eigen bevinding uit de derde studie dat een procedure met 2 kritische grenzen tot stabiele resultaten kan leiden. Via de toevoeging van de gemiddelde herselectieratio's van de voxels in een cluster werd sta-

biliteit opgenomen als beslissingscriterium. Zonder een verlies op andere vlakken resulteert dit in de geïnformeerde selectie van meer clusters.

In de algemene discussie sloten we dit proefschrift af door 3 centrale onderwerpen dieper te behandelen. Een eerste onderwerp is deze keuze van inferentiestrategie (stap 4). In studie 1-3 was dit een recurrent thema. Op basis van onze studies vinden we dat de resultaten die bekomen worden op basis van zogenaamde niet-parametrische technieken een waardig alternatief kunnen bieden voor de klassieke inferentiestrategieën. Zeker omdat bij die klassieke strategieën vaak de assumpties, een noodzakelijke voorwaarde voor een valide toepassing, niet gecontroleerd worden ondanks de beschikbaarheid van technieken (Zhang, Luo, & Nichols, 2006).

Een tweede belangrijk onderwerp is de introductie van data-analytische stabiliteit binnen de evaluatie van methoden in fMRI. De toevoeging van de stabiliteit bracht bovendien extra inzicht in de impact van methodologische keuzes. Bovendien kan ook een van data-analytische stabiliteit afgeleide maat ook opgenomen worden in de beslissing of een cluster/voxel actief is. Dit is het derde thema uit de discussie. We vonden dat dit op een intuïtieve manier kan toegevoegd worden in de beslissingscontext.

Conclusies In dit proefschrift hebben we het concept van data-analytische stabiliteit vertaald naar de analyse van fMRI data. Hiermee kunnen we de reproduceerbaarheid van een analyse kwantificeren binnen een replicatiecontext. In de eerste studie onderzochten we de kwaliteiten voor het opzetten van deze replicatiecontext. In de tweede en derde studie hebben we data-analytische stabiliteit gebruikt om de impact van methodologische keuzes te onderzoeken. En in de vierde studie werd stabliteit toegevoegd als criterium in de beslissing over activatie in een regio. Met proefschrift hopen we op een kwantificeerbare manier reproduceerbaarheid onder de aandacht gebracht te hebben in de complexe analyse van fMRI data.

Bibliografie

Bennett, C. M., & Miller, M. B. (2013). fmri reliability: Influences of task and experimental design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690–702.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300.

Eklund, A., Nichols, T., & Knutsson, H. (2015). Can parametric statistical methods be trusted for fMRI based group studies? *ArXiv e-prints*.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, *7*.

Zhang, H., Luo, W.-L., & Nichols, T. E. (2006). Diagnosis of single-subject and group fMRI data with SPMd. *Human brain mapping*, *27*(5), 442–51.

8 Data Storage Fact Sheets

Data Storage Fact Sheet Chapter 2

```
 % Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Sanne Roels, Chapter 2.
% Author: Sanne Roels
% Date: 10/03/2016


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Sanne Roels
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: sanne.roels@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Beatrijs Moerkerke
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: beatrijs.moerkerke@ugent.be

If a response is not received when using the above contact details,
     please send an email to data.pp@ugent.be or contact Data
     Management, Faculty of Psychology and Educational Sciences, Henri
     Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Roels, S. P., Moerkerke, B., \& Loeys, T. (2015). Bootstrapping fMRI
     Data: Dealing with Misspecification. \emph{NeuroInformatics, 13,}
     337  352  .

* Which datasets in that publication does this sheet apply to?:
Scripts for data generation + raw data from example in paper


3. Information about the files that have been stored
============================================================


3a. Raw data
```

```
------------------------------------------------------------

* Have the raw data been stored by the main researcher? [X] YES / [ ]
    NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [ ] research group file server
  - [X] other (specify): hard disk responsible ZAP

* Who has direct access to the raw data (i.e., without intervention of
    another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
      results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: Scripts to genererate the
      data, scripts to analyze generated and raw data.
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
      this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: hard disk responsible ZAP

* Who has direct access to these other files (i.e., without
    intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
```

```
   - e-mail:

v0.2
```

Data Storage Fact Sheet Chapter 3

```
 % Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Sanne Roels, Chapter 3.
% Author: Sanne Roels
% Date: 10/03/2016


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Sanne Roels
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: sanne.roels@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Beatrijs Moerkerke
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: beatrijs.moerkerke@ugent.be

If a response is not received when using the above contact details,
     please send an email to data.pp@ugent.be or contact Data
     Management, Faculty of Psychology and Educational Sciences, Henri
     Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Roels, S. P., Bossier, H., Loeys, T., \& Moerkerke, B. (2015). Data-
     analytical stability of cluster-wise and peak-wise inference in
     fMRI data analysis. \emph{Journal of Neuroscience Methods, 240},
     37  47  .

* Which datasets in that publication does this sheet apply to?:
Scripts for data generation + raw data from example in paper


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher? [X] YES / [ ]
     NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [ ] research group file server
  - [X] other (specify): hard disk responsible ZAP
```

```
* Who has direct access to the raw data (i.e., without intervention of
     another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


3b. Other files
---------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
       results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: Scripts to generate the
       data, scripts to analyze generated and raw data.
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
       this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: hard disk responsible ZAP

* Who has direct access to these other files (i.e., without
     intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
===============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```

Data Storage Fact Sheet Chapter 4

```
 % Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Sanne Roels, Chapter 4.
% Author: Sanne Roels
% Date: 10/03/2016


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Sanne Roels
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: sanne.roels@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Beatrijs Moerkerke
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: beatrijs.moerkerke@ugent.be

If a response is not received when using the above contact details,
     please send an email to data.pp@ugent.be or contact Data
     Management, Faculty of Psychology and Educational Sciences, Henri
     Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Roels, S. P., Loeys, T., \& Moerkerke, B. (2016). Evaluation of Second-
     Level Inference in fMRI Analysis. \emph{Computational Intelligence
      and Neuroscience, 2016}, Article ID 1068434, 22 pages.

* Which datasets in that publication does this sheet apply to?:
Scripts for data generation and analysis + raw data from example in
     paper


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher? [] YES / [X] NO
If NO, please justify:
Data come from the Human Connectome Project. This data set is freely
     available on: http://www.humanconnectome.org/data/


* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
```

```
  - [X] other (specify): publicly available data: http://www.
      humanconnectome.org/data/


* Who has direct access to the raw data (i.e., without intervention of
    another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [X] all members of UGent
  - [X] other (specify): publicly available data via http://www.
      humanconnectome.org/data/


3b. Other files
---------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
      results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: Scripts to generate the
      data, scripts to analyze generated and raw data.
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
      this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: hard disk responsible ZAP

* Who has direct access to these other files (i.e., without
    intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
===============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```

Data Storage Fact Sheet Chapter 5

```
 % Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Sanne Roels, Chapter 5.
% Author: Sanne Roels
% Date: 10/03/2016


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Sanne Roels
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: sanne.roels@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Beatrijs Moerkerke
- address: H. Dunantlaan 1, 9000 Gent
- e-mail: beatrijs.moerkerke@ugent.be

If a response is not received when using the above contact details,
     please send an email to data.pp@ugent.be or contact Data
     Management, Faculty of Psychology and Educational Sciences, Henri
     Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
At the time of the submission of the PhD dissertation this chapter was
     not published.

* Which datasets in that publication does this sheet apply to?:
Scripts for the analysis


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher? [] YES / [X] NO
If NO, please justify:
Data come from the Human Connectome Project. This data set is freely
     available on: http://www.humanconnectome.org/data/

* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
  - [X] other (specify): publicly available data: http://www.
       humanconnectome.org/data/
```

```
* Who has direct access to the raw data (i.e., without intervention of
     another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [X] all members of UGent
  - [X] other (specify): publicly available data via http://www.
       humanconnectome.org/data/


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
       results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: Scripts to analyze the
       data
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
       this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: hard disk responsible ZAP

* Who has direct access to these other files (i.e., without
     intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```